Introduction to Artificial Intelligence

丁尧相
浙江大学

Fall & Winter 2022
Week 7

Announcements

- We will release Problem Set 2.2 after this lecture.
- We will release some more materials about lab project 1.
- The mid-term quiz will be on next Monday. If you can't participate on class, please let me know before this Wednesday.

Knowledge Reasoning: III

- Probabilistic Reasoning
  - Bayes net: Approximate inference
- Causal reasoning
- Take-home messages

Knowledge Reasoning: III

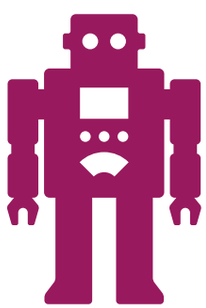
- Probabilistic Reasoning
  - Bayes net: Approximate inference
- Causal reasoning
- Take-home messages

Internal vs. External Model

Since the agent cannot fully know the external model,
it should build an internal model itself for decision making.

internal model

agent



partial state

utility

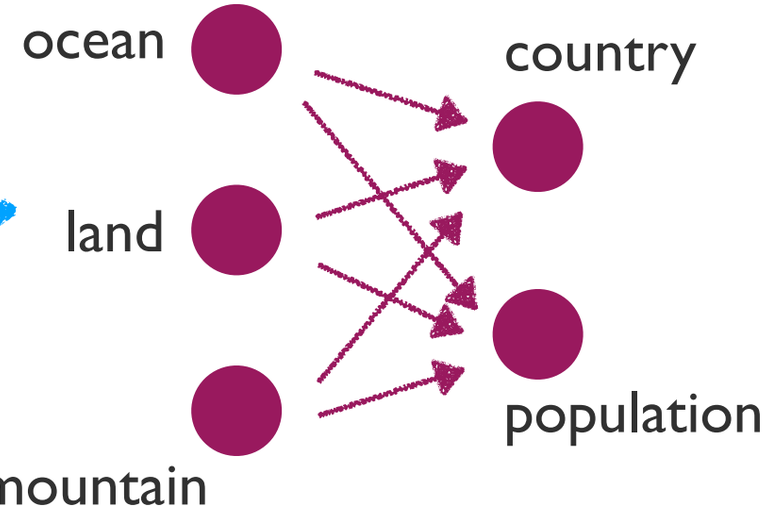
action

environment
external model

Knowledge in AI Systems

longitude and latitude

ocean, land, mountain

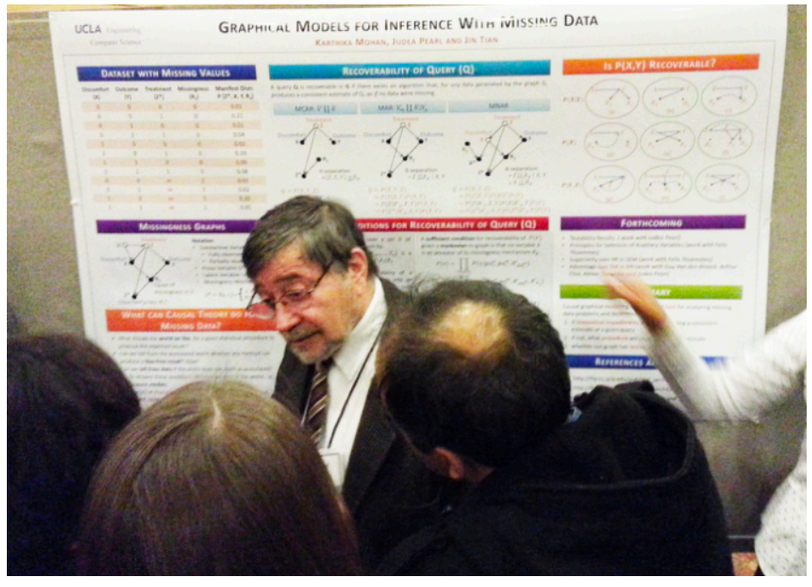


- Turn primitive external states into meaningful internal states.
- Reason about most useful states for decision making.
- Capture internal relationships among factors of decision making.

These reasoning rules are called knowledges in an AI system.

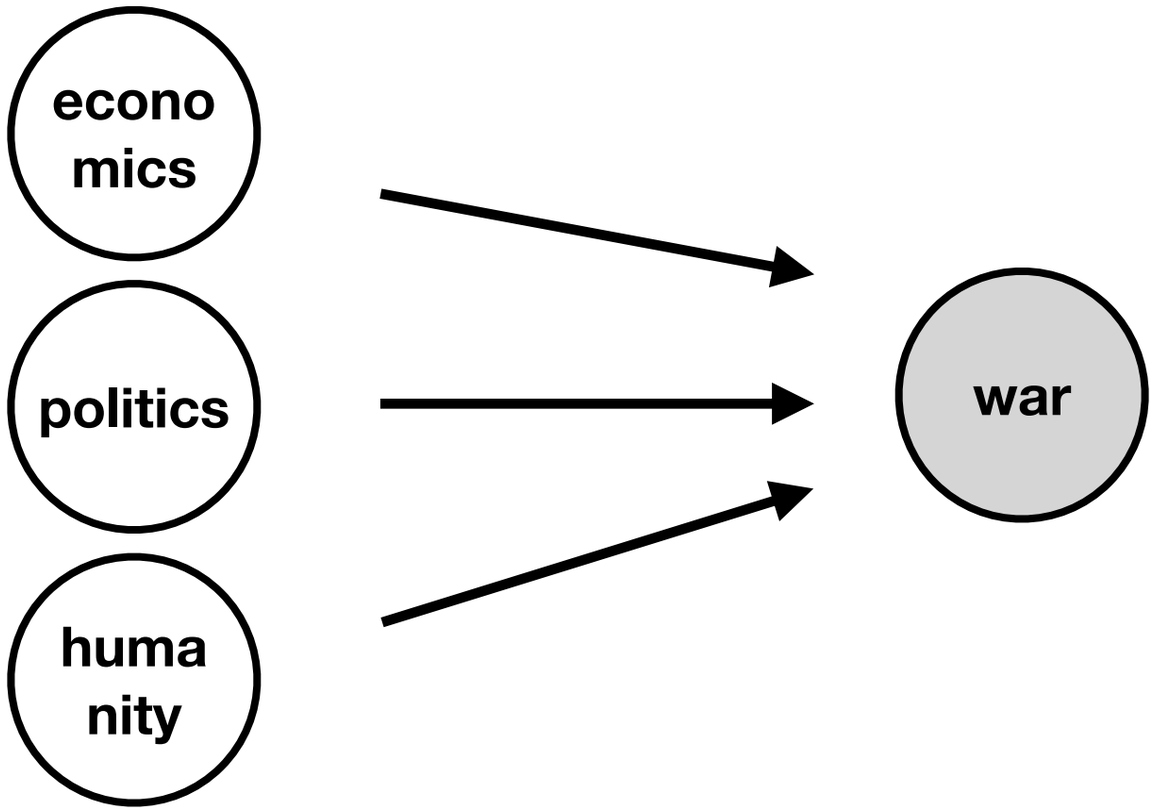
From Deterministic to Probabilistic Reasoning

- Modeling human knowledge, e.g. common sense:
  - Uncertainty in reasoning
  - Fast learning and processing
- We will introduce two foundational tools:
  - Bayes nets
  - Causal reasoning



Judea Pearl

Bayes Nets



- Bayes nets are acyclic directed graphs representing the joint distribution of random variables.

- The nodes represent random variables, the edges represent the relationship among r. v. (e.g. hidden to observable r. v.)

Basic Tasks in Probabilistic Reasoning

- In probabilistic reasoning, we try to model the joint distribution of a set of random variables $P(X_1, X_2, ..., X_n)$ and do:
  - Inference: answering queries about the marginal distributions.
  - Conditional independence test: decide the conditional independence of a subset of random variables.
  - Learning: obtain the structure of the joint distribution.

Inference is to reasoning about the value of the variables.
The independence test and learning are to understand relationship among variables.

The Inference Problem

- Inference in probabilistic models: calculate margin quantities from the joint distribution represented by the model.

- Examples:
  - Posterior probability

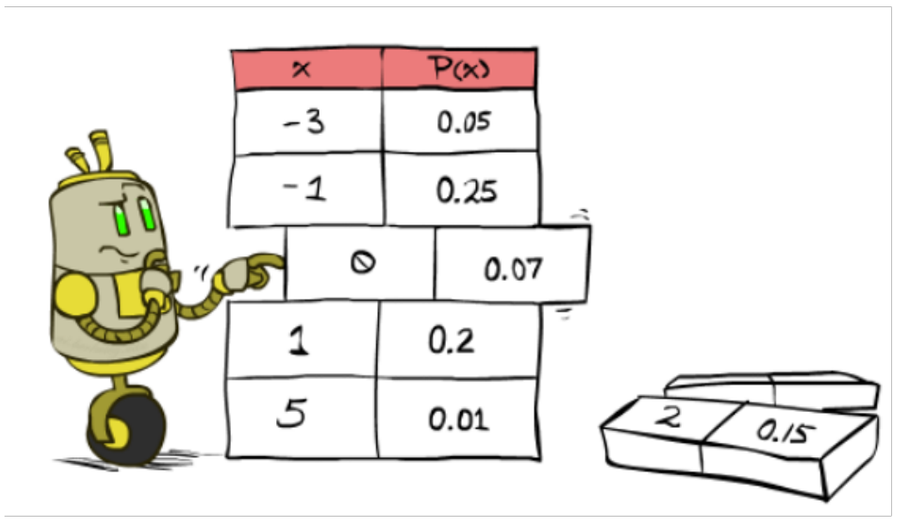
$$P(Q|E_1 = e_1, \ldots E_k = e_k)$$

  - Most likely explanation:

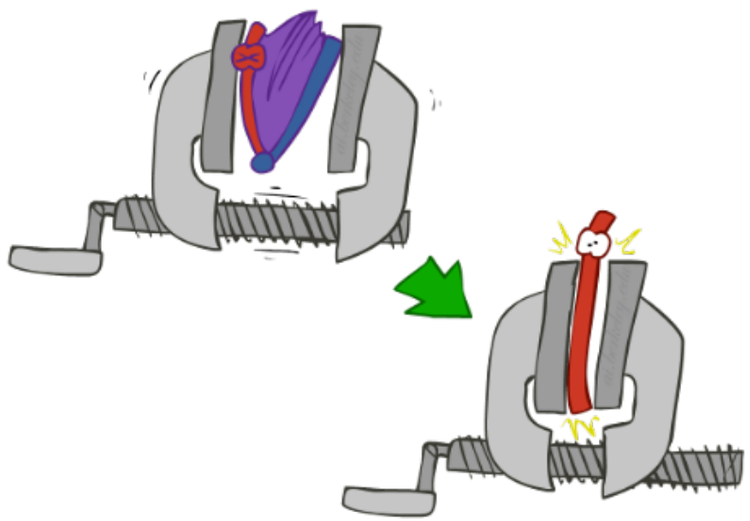
$$\text{argmax}_q \; P(Q = q|E_1 = e_1 \ldots)$$

Inference by Enumeration

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



- Step 3: Normalize

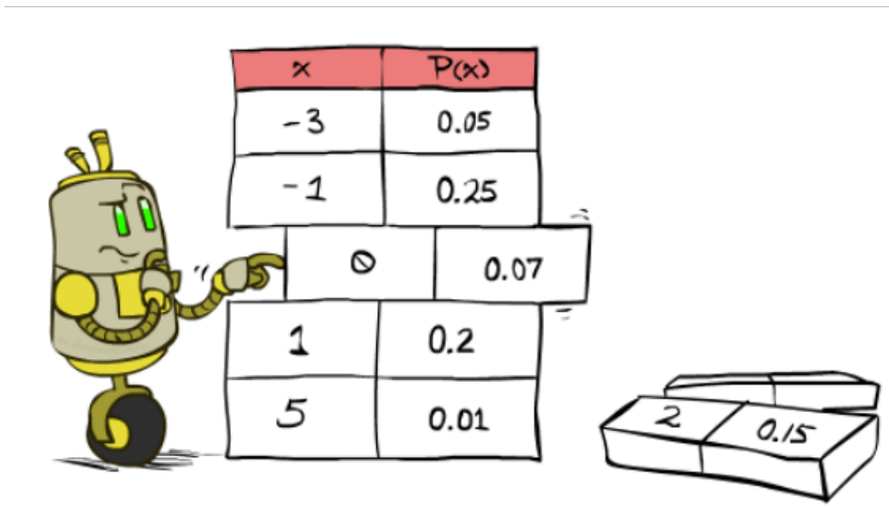
$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

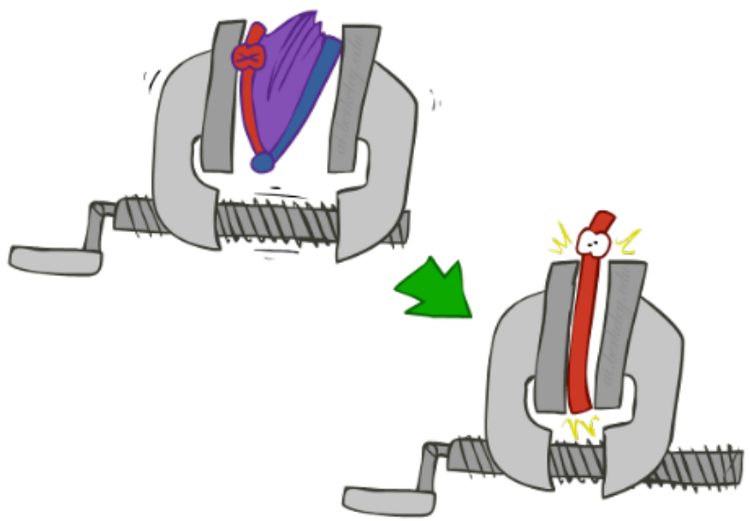
Slide courtesy: Dan Klein & Pieter Abbeel

Inference by Enumeration

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

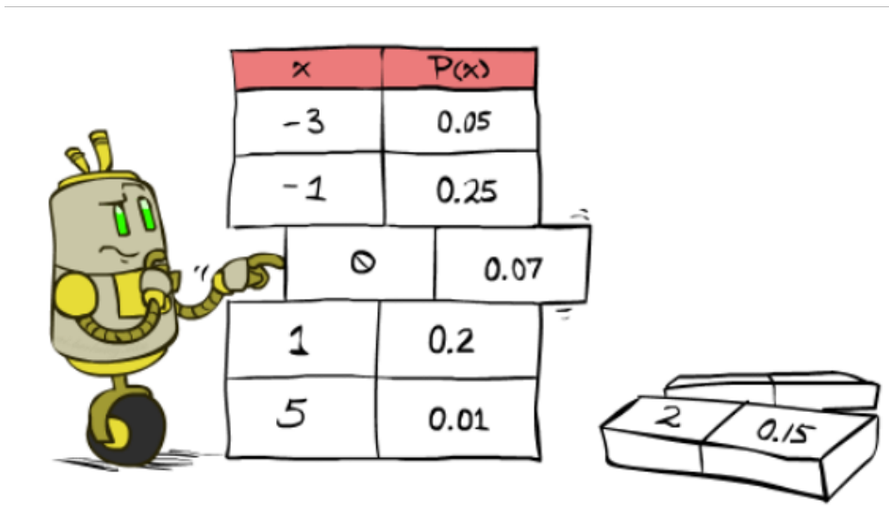
$$P(war = true | hum = kind)$$

$$= \sum_{i,j} p(war = true, eco = i, pol = j, hum = kind)/Z$$

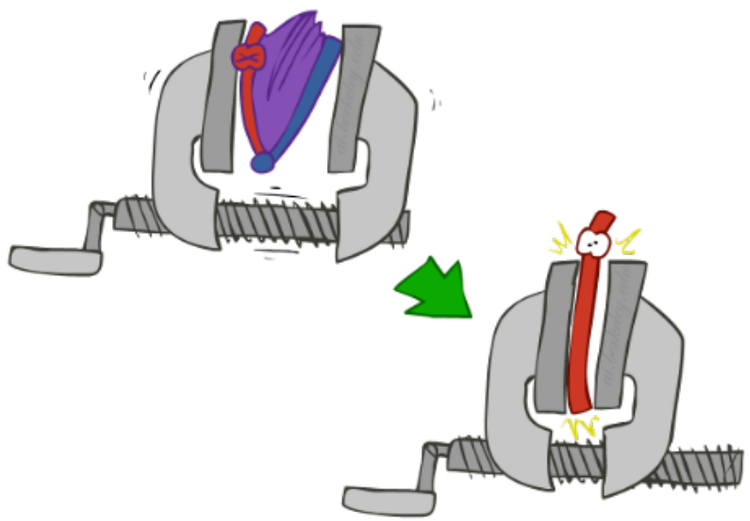
Slide courtesy: Dan Klein & Pieter Abbeel

Inference by Enumeration

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

$$P(war = true | hum = kind)$$

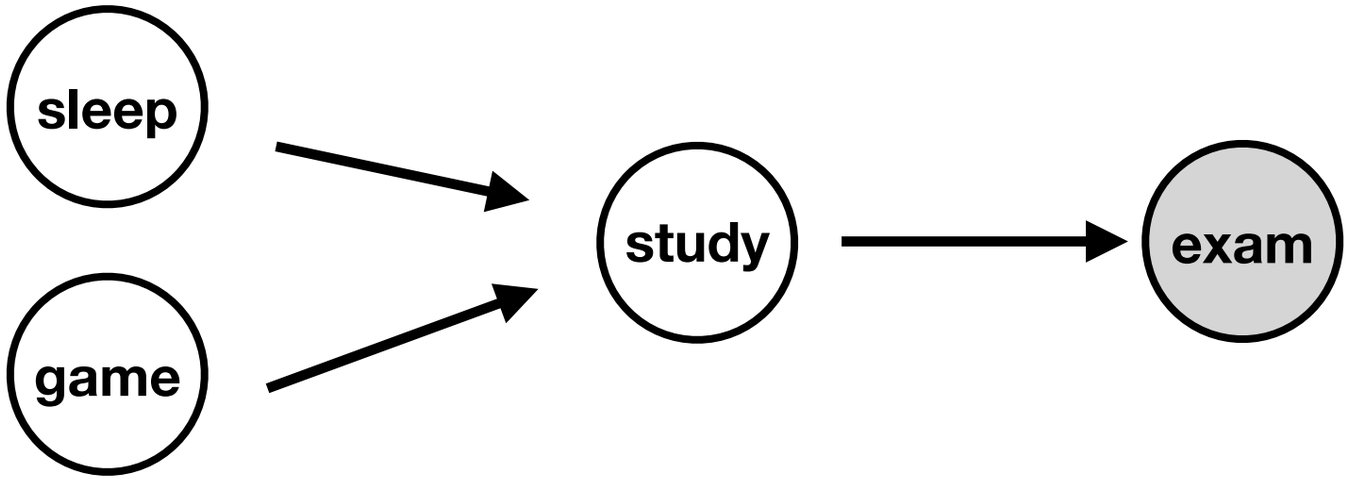
$$= \sum_{i,j} p(war = true, eco = i, pol = j, hum = kind)/Z$$

$$Z = p(hum = kind)$$

$$= \sum_{i,j,k} p(war = k, eco = i, pol = j, hum = kind)$$

Slide courtesy: Dan Klein & Pieter Abbeel

Variable Elimination



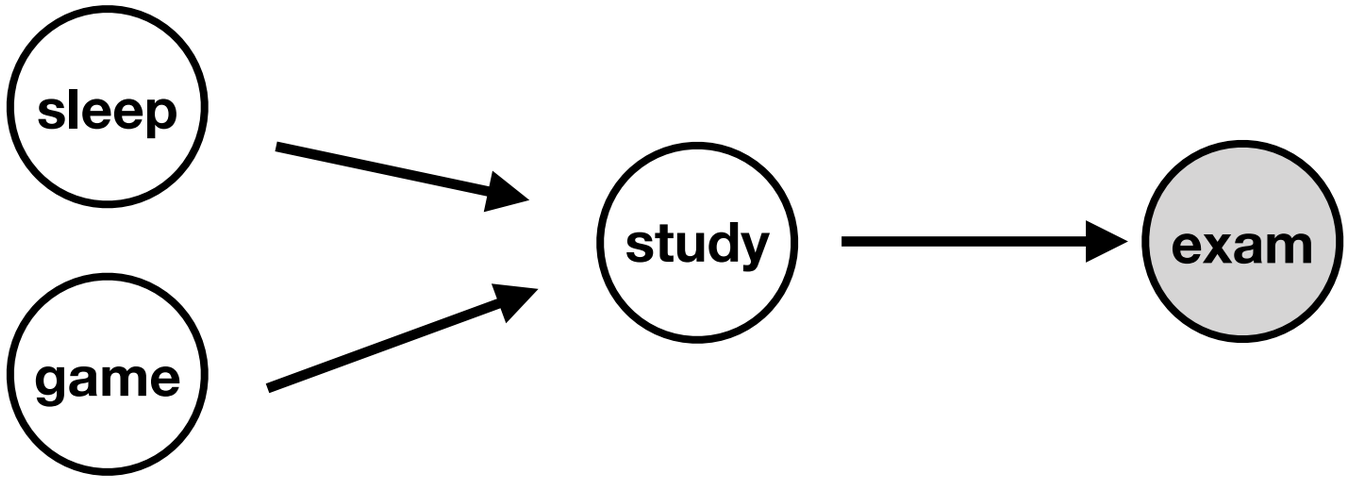
- The key idea is to marginalize early by utilizing the structure of BN.

- To answer: $p(exam|sleep = true)$

- Enumeration:

$$\sum_{i,j} p(exam|study = i)p(study = i|sleep = true, game = j)p(sleep = true)p(game = j)$$

Variable Elimination



- The key idea is to marginalize early by utilizing the structure of BN.

- To answer: $p(exam|sleep = true)$

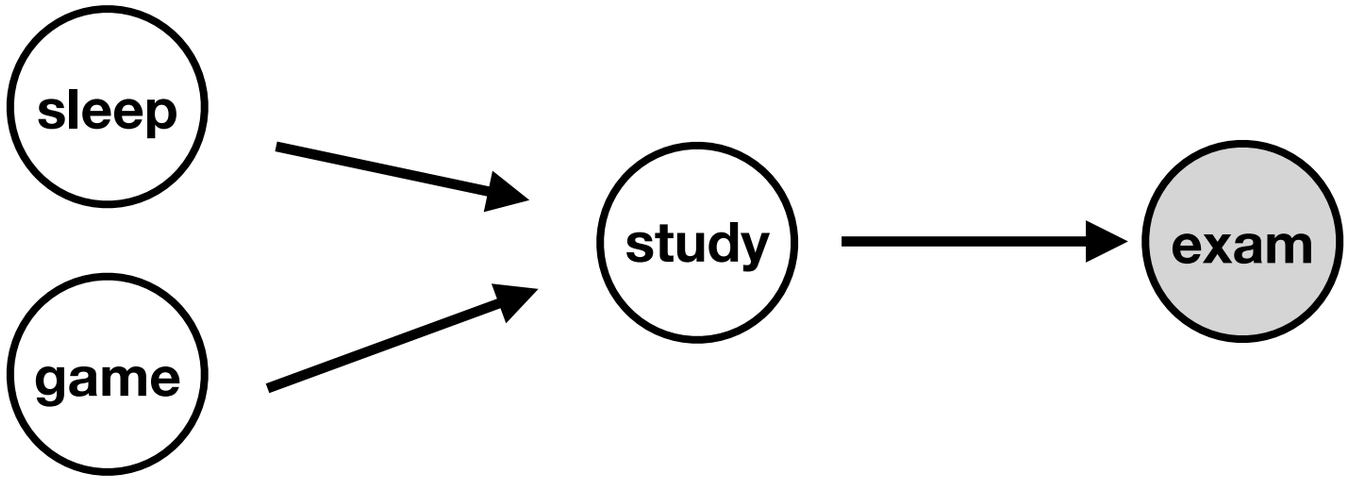
- Enumeration:

$$\sum_{i,j} p(exam|study = i)p(study = i|sleep = true, game = j)p(sleep = true)p(game = j)$$

- Elimination:

$$\sum_{i} p(sleep = true)p(exam|study = i) \sum_{j} p(study = i|sleep = true, game = j)p(game = j)$$

Variable Elimination



- The key idea is to marginalize early by utilizing the structure of BN.

- To answer: $p(exam|sleep = true)$

- Enumeration:

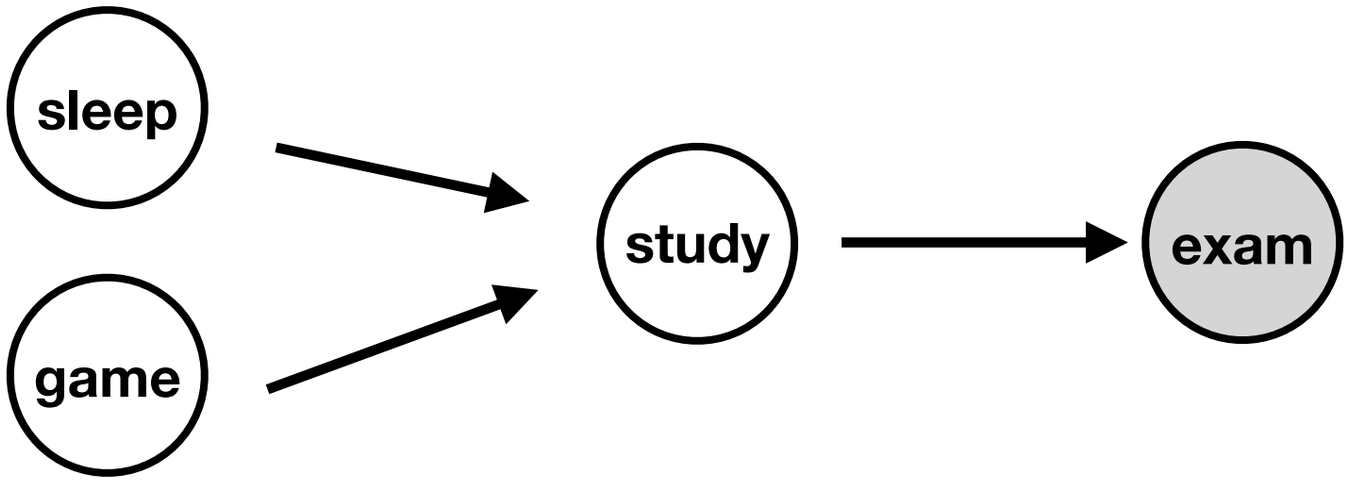
$$\sum_{i,j} p(exam|study = i)p(study = i|sleep = true, game = j)p(sleep = true)p(game = j)$$

- Elimination:

Sum out hidden variables early!

$$\sum_{i} p(sleep = true)p(exam|study = i) \sum_{j} p(study = i|sleep = true, game = j)p(game = j)$$

Variable Elimination



- The key idea is to marginalize early by utilizing the structure of BN.

- To answer: $p(exam|sleep = true)$

- Enumeration:

$$\sum_{i,j} p(exam|study = i)p(study = i|sleep = true, game = j)p(sleep = true)p(game = j)$$

- Elimination:

Sum out hidden variables early!

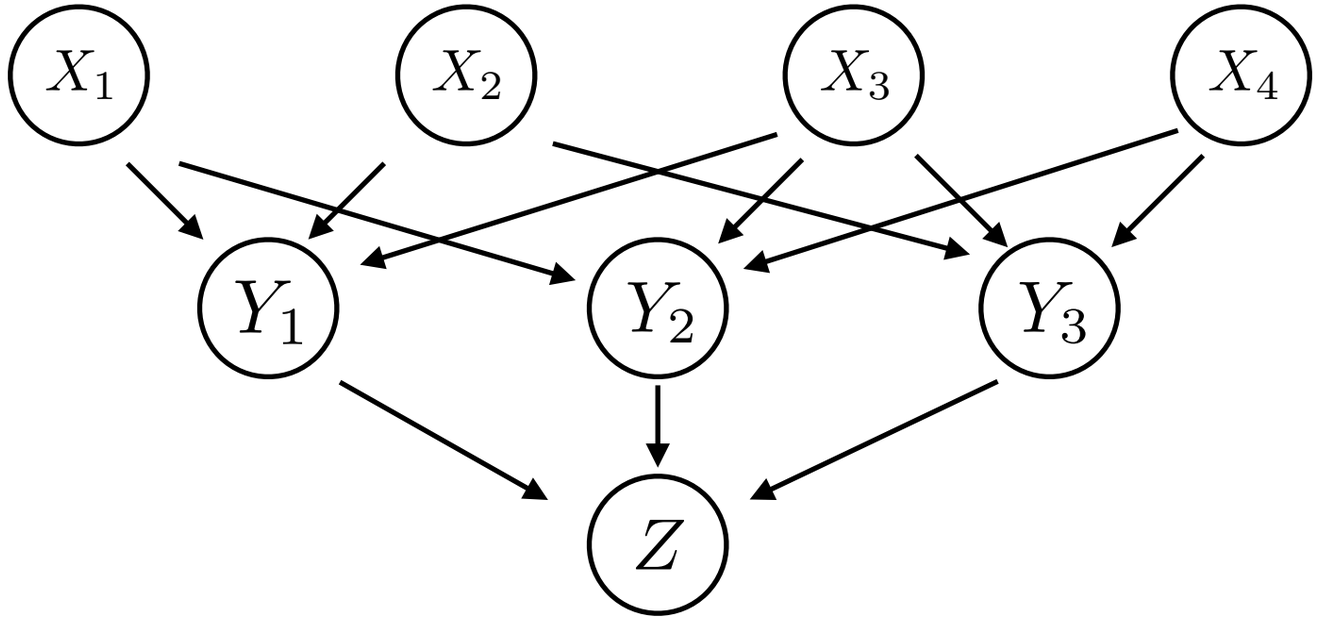
$$\sum_{i} p(sleep = true)p(exam|study = i) \sum_{j} p(study = i|sleep = true, game = j)p(game = j)$$

Can save much by avoiding later enumeration!

Computational Complexity

- Is variable elimination a poly-time algorithm?
- Consider representing 3-CNF as BN:

$$(X_1 \vee X_2 \vee \vee \neg X_3) \wedge (\neg X_1 \vee X_3 \vee X_4) \wedge (X_2 \vee \neg X_3 \vee X_4)$$



If we can determine whether $p(Z) = 0$, then we can solve the 3-SAT problem which is NP-complete. So the exact inference in BN is NP-hard.

Knowledge Reasoning: III

- Probabilistic Reasoning
  - Bayes net: Approximate inference
- Causal reasoning
- Take-home messages

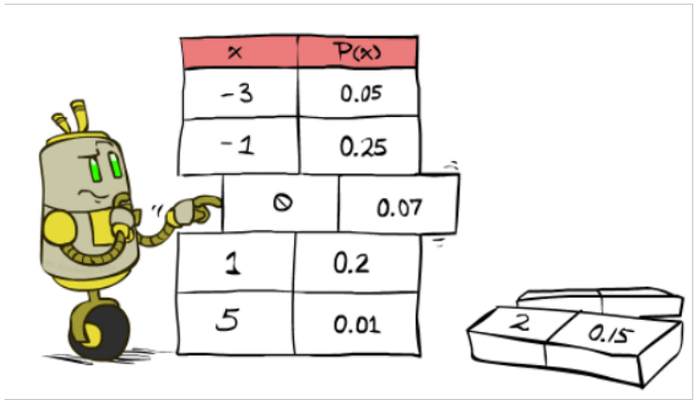
Why Inference is Hard?

- Target: estimate the conditional probability:

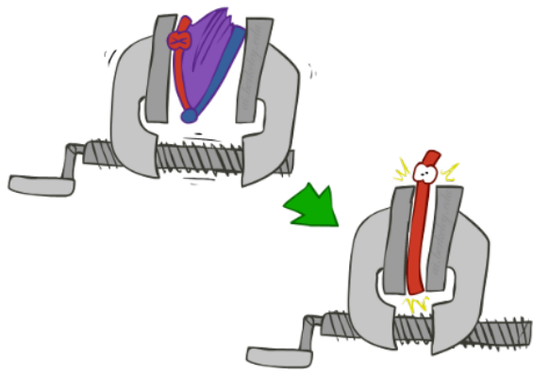
$$P(Q|E_1 = e_1, ...E_k = e_k) = P(Q, E_1 = e_1, ..., E_k = e_k)/P(E_1 = e_1, ..., E_k = e_k)$$

- The summations in marginalization cost most.

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



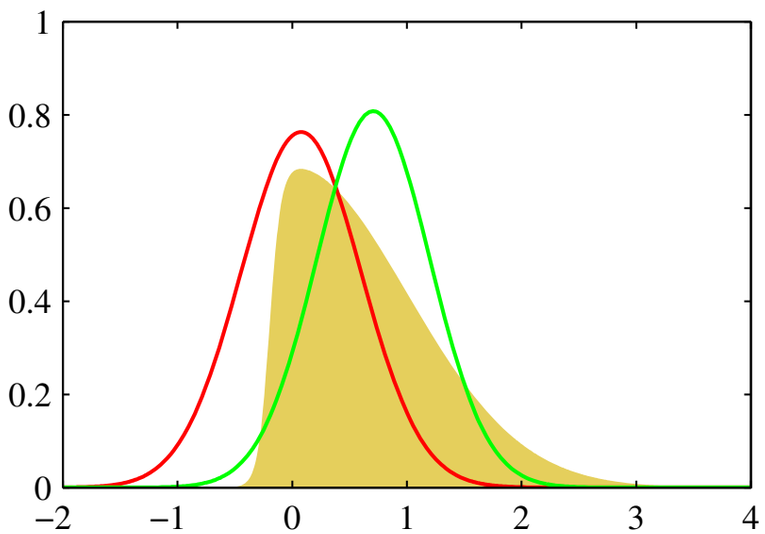
- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

Approximation Inference Approaches

- Estimating $P(Q|E_1 = e_1, ...E_k = e_k)$
- Variational Inference
  - Using simpler $P'(Q|E_1 = e_1, ..., E_k = e_k)$ for approximation
- Sampling
  - Empirical estimation based on samples (similar to MCTS)



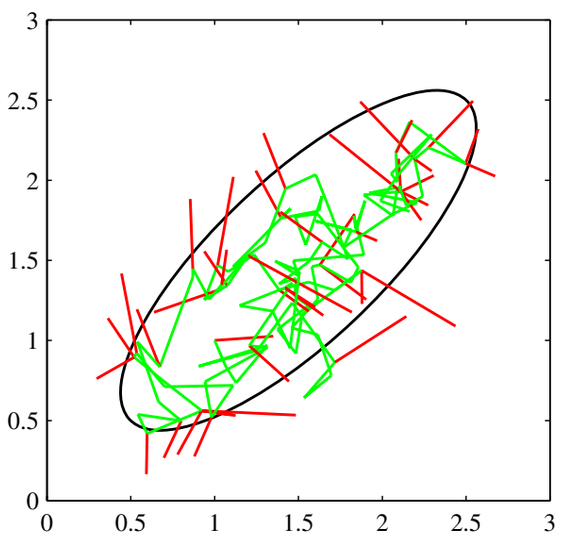
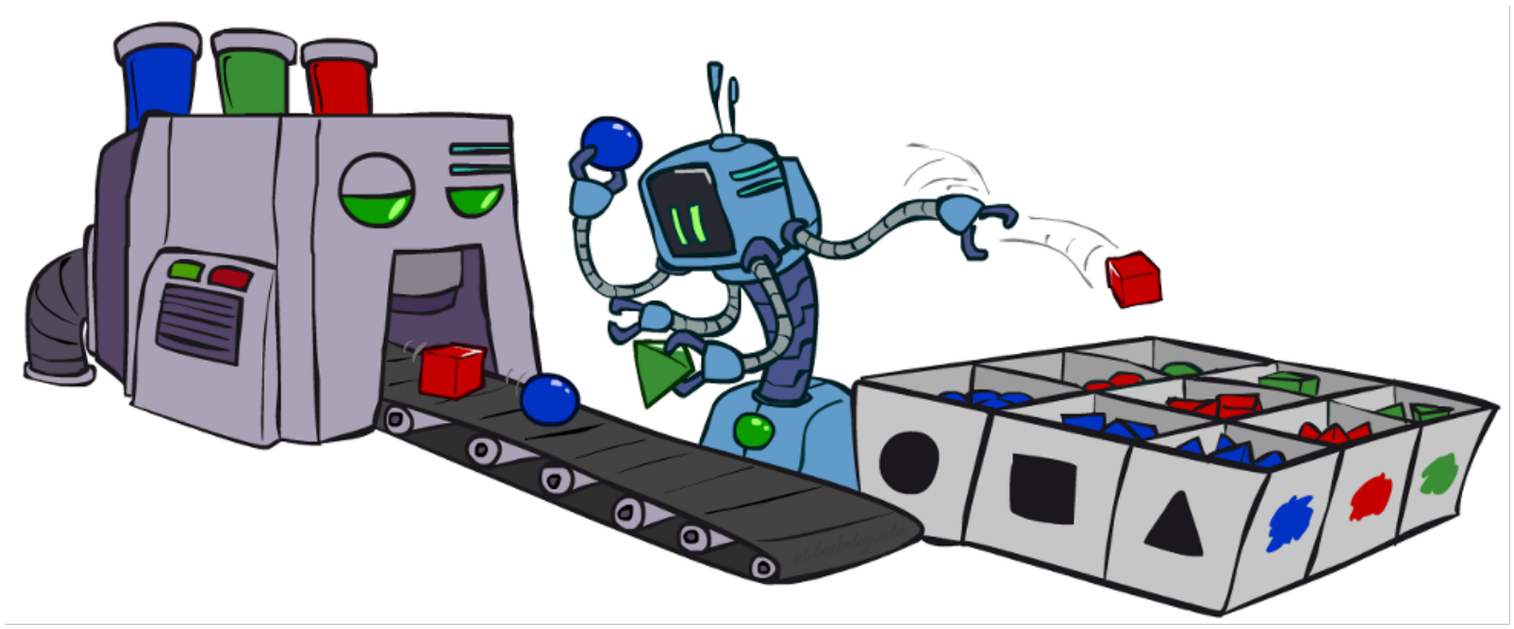


Figure Courtesy:
Christopher M. Bishop

Approximation by Sampling

- Basic idea
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability
  - Show this converges to the true probability P



Slide courtesy: Dan Klein & Pieter Abbeel

Sampling Methods

- Prior sampling
- Rejection sampling
- Likelihood weighting
- Gibbs sampling

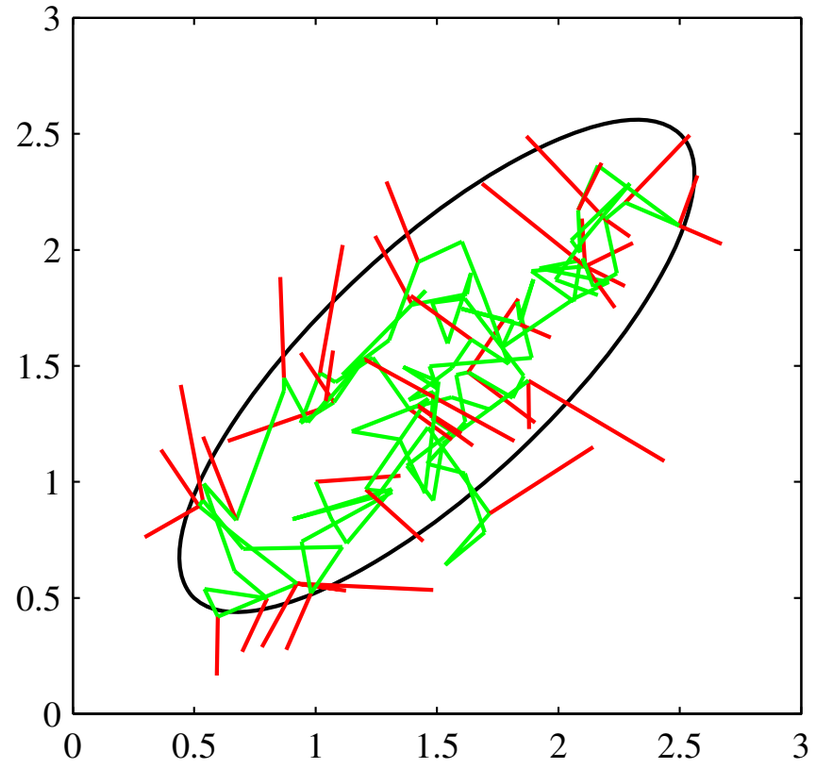


Figure Courtesy: Christopher M. Bishop

Prior Sampling

- Suppose the target is to estimate the joint distribution

$$S_{PS}(x_1 \dots x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(X_i)) = P(x_1 \dots x_n)$$

- The key idea is to generate samples of $x_1, x_2, \ldots, x_n$ w.r.t. the probabilities $P(x_i|Parents(X_i))$

- The number of samples: $N_{PS}(x_1 \dots x_n)$

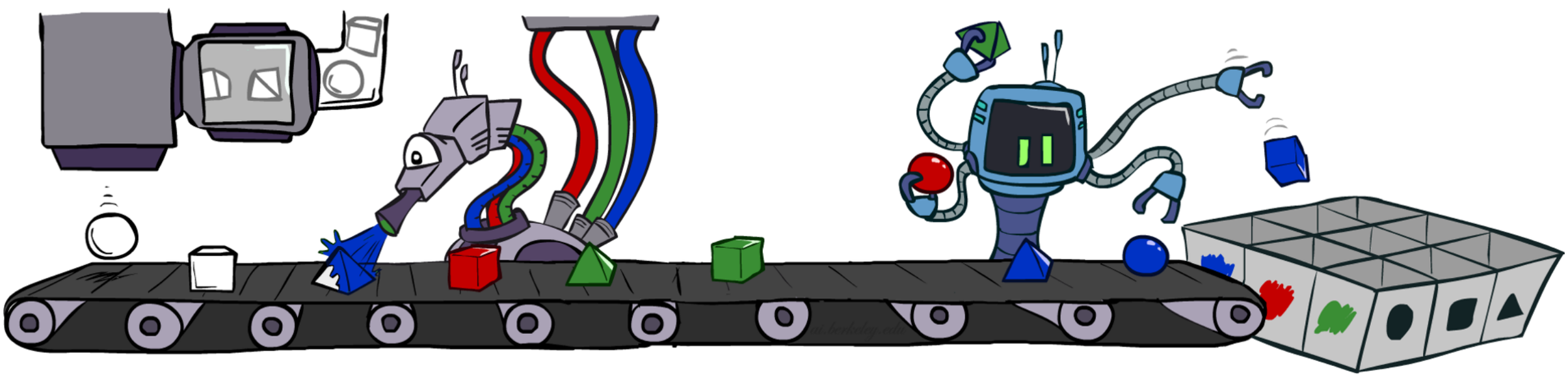
- Consistency:

$$\begin{aligned} \lim_{N\to\infty} \hat{P}(x_1, \dots, x_n) &= \lim_{N\to\infty} N_{PS}(x_1, \dots, x_n)/N \\ &= S_{PS}(x_1, \dots, x_n) \\ &= P(x_1 \dots x_n) \end{aligned}$$

Slide courtesy: Dan Klein & Pieter Abbeel

Prior Sampling

- For i=1, 2, …, n
  - Sample x_i from $P(X_i \mid Parents(X_i))$
- Return $(x_1, x_2, \ldots, x_n)$



Slide courtesy: Dan Klein & Pieter Abbeel

Prior Sampling

- Generate samples:

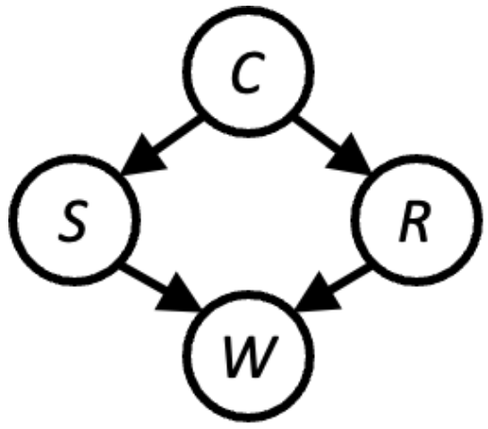
+c, -s, +r, +w

+c, +s, +r, +w

-c, +s, +r, -w

+c, -s, +r, +w

-c, -s, -r, +w



- About $P(c)$
  - $\hat{P}(+c) = 3/5, \hat{P}(-c) = 2/5.$

- How to calculate $P(c|W), P(c|S, R)$

Slide courtesy: Dan Klein & Pieter Abbeel

Prior Sampling

- Generate samples:

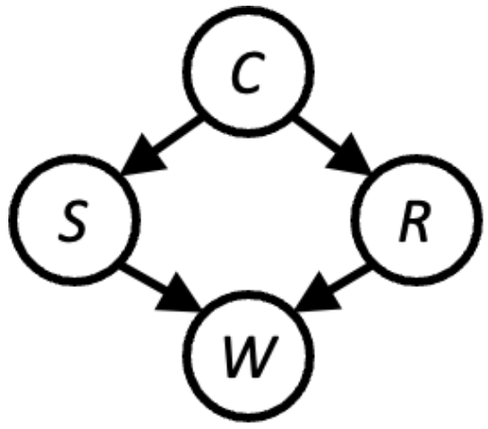
+c, -s, +r, +w

+c, +s, +r, +w

-c, +s, +r, -w

+c, -s, +r, +w

-c, -s, -r, +w



- About $P(c)$
  - $\hat{P}(+c) = 3/5, \hat{P}(-c) = 2/5.$

- How to calculate $P(c|W), P(c|S, R)$

How to calculate probabilities conditioned on evidences?

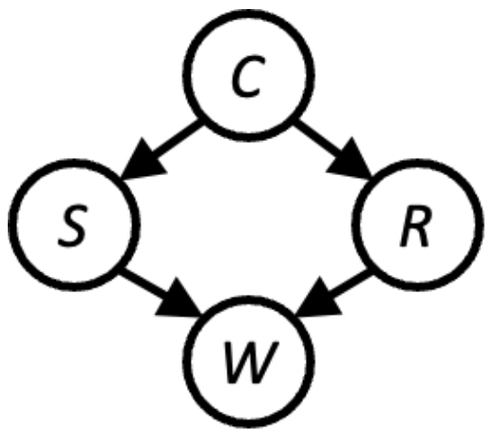
Slide courtesy: Dan Klein & Pieter Abbeel

Rejection Sampling

- Suppose the target is to estimate the distribution conditioned on evidence $P(C|+S)$

- The key idea is to generate samples w.r.t. the graph, and then reject the samples without $+S$

- consistency property also holds

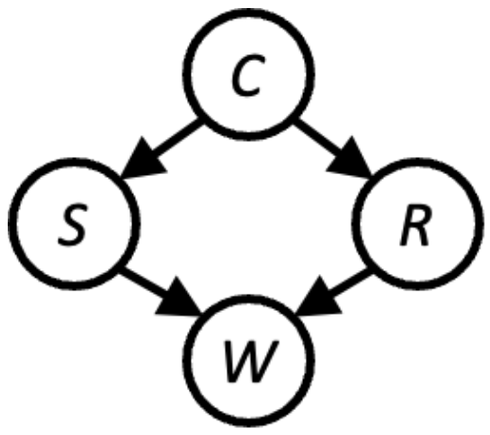


Rejection Sampling

- Suppose the target is to estimate the distribution conditioned on evidence $P(C|+S)$

- The key idea is to generate samples w.r.t. the graph, and then reject the samples without $+S$

- consistency property also holds



+c, -s, +r, +w

+c, +s, +r, +w

-c, +s, +r, -w

+c, -s, +r, +w

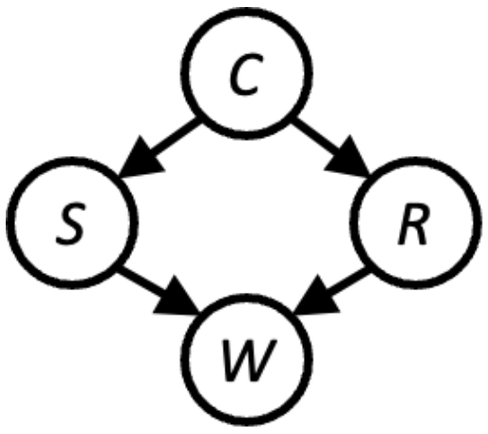
-c, -s, -r, +w

Rejection Sampling

- Suppose the target is to estimate the distribution conditioned on evidence $P(C|+S)$

- The key idea is to generate samples w.r.t. the graph, and then reject the samples without $+S$

- consistency property also holds



+c, -s, +r, +w

+c, +s, +r, +w

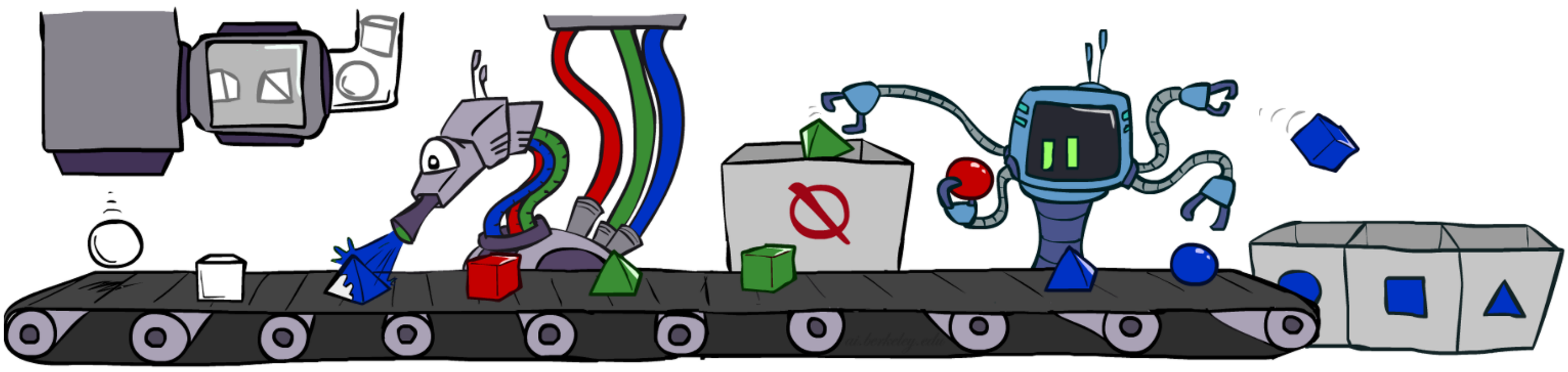
-c, +s, +r, -w

+c, -s, +r, +w

-c, -s, -r, +w

Rejection Sampling

- IN: evidence instantiation
- For i=1, 2, …, n
  - Sample x_i from $P(X_i \mid \text{Parents}(X_i))$
  - If x_i not consistent with evidence
    - Reject: Return, and no sample is generated in this cycle
- Return $(x_1, x_2, \ldots, x_n)$



Slide courtesy: Dan Klein & Pieter Abbeel

Rejection Sampling

- Generate samples:

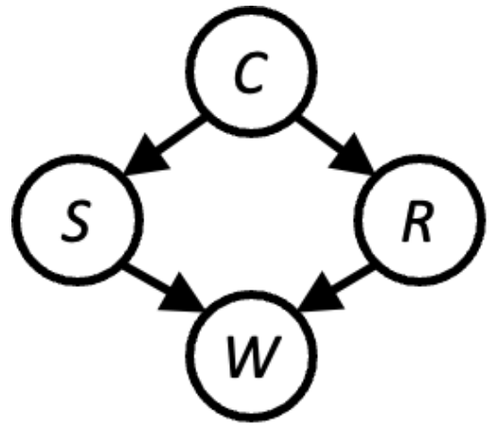
+c, -s, +r, +w

+c, +s, +r, +w

-c, +s, +r, -w

+c, -s, +r, +w

-c, -s, -r, +w



- About $P(C|+S)$
  - $\hat{P}(+c|+S) = \sout{2/3}$, $\hat{P}(-c|+S) = \sout{1/3}$.

1/2 **1/2**

- Any drawback exists?

Rejection Sampling

- Generate samples:

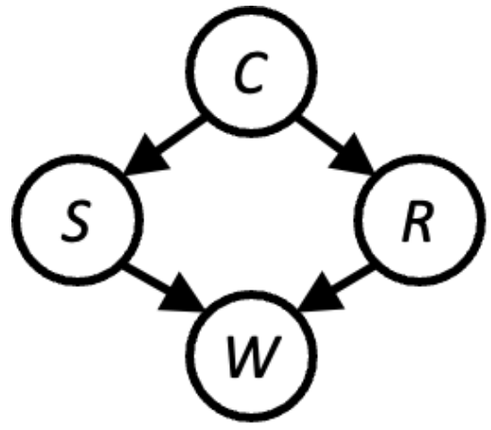
+c, -s, +r, +w

+c, +s, +r, +w

-c, +s, +r, -w

+c, -s, +r, +w

-c, -s, -r, +w

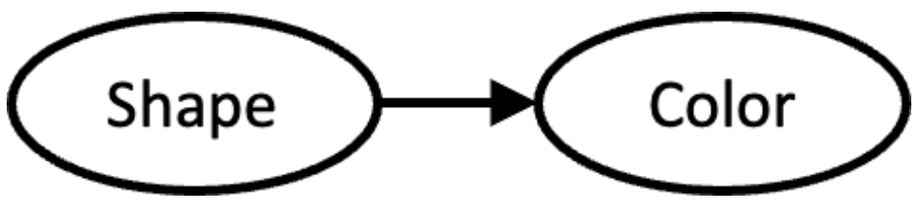


- About $P(C|+S)$
  - $\hat{P}(+c|+S) = \sout{2/3}\ \mathbf{1/2}, \hat{P}(-c|+S) = \sout{1/3}\ \mathbf{1/2}$
- Any drawback exists?

The samples are mostly wasteful when the evidences rarely happen!
Does not use the evidences when generating samples!

Likelihood Weighting

- Consider $P(shape|color = blue)$

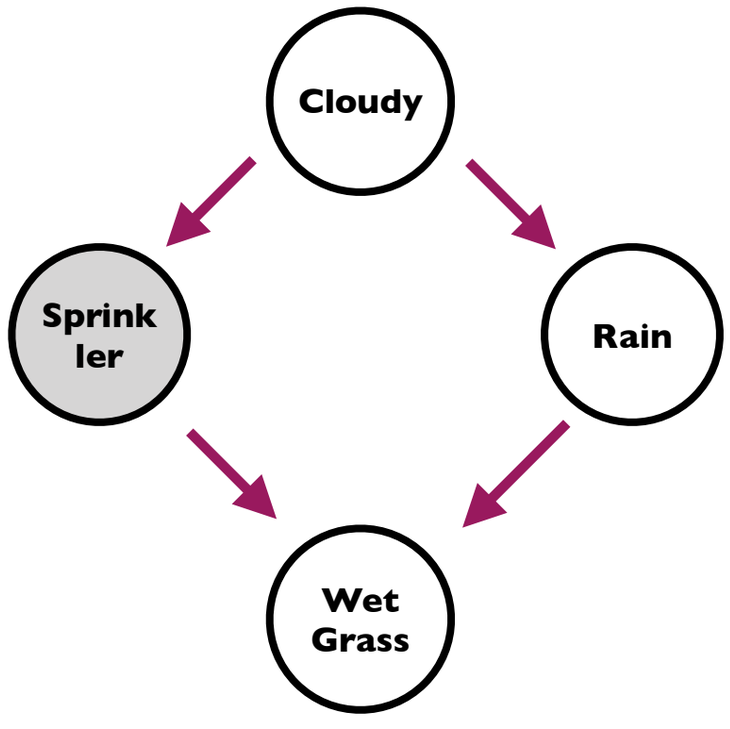


- Rejection sampling wastes lots of samples:

~~pyramid, green~~
~~pyramid, red~~
sphere, blue
~~cube, red~~
~~sphere, green~~

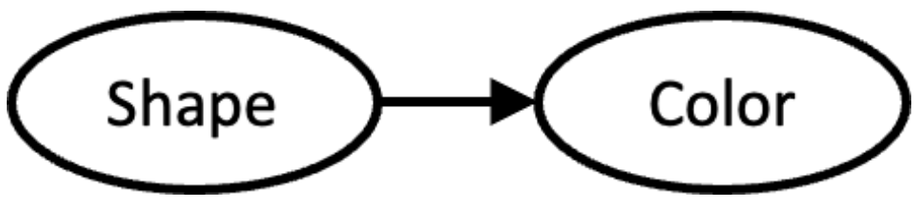
pyramid, blue
pyramid, blue
sphere, blue
cube, blue
sphere, blue

- Why not generate samples directly with conditioning on the evidence?



Likelihood Weighting

- Consider $P(shape|color = blue)$



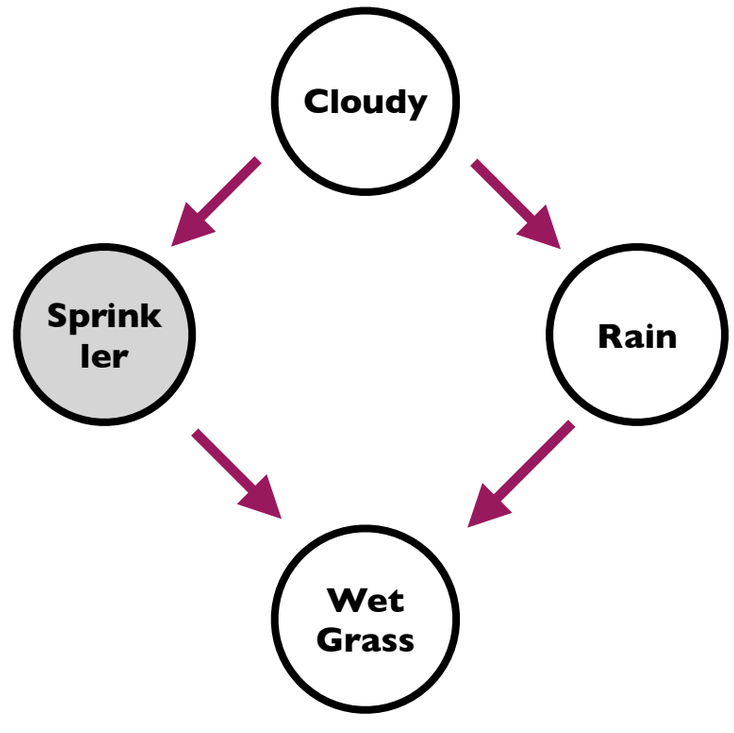
- Rejection sampling wastes lots of samples:

~~pyramid, green~~
~~pyramid, red~~
sphere, blue
~~cube, red~~
~~sphere, green~~

pyramid, blue
pyramid, blue
sphere, blue
cube, blue
sphere, blue

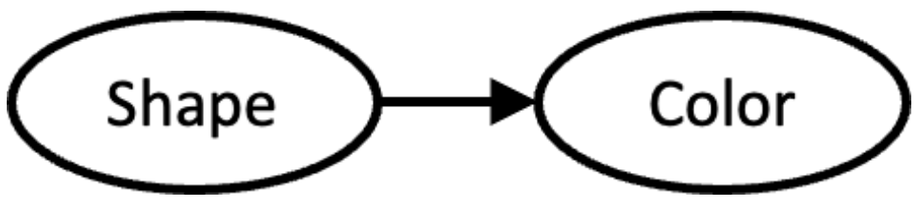
- Why not generate samples directly with conditioning on the evidence?

The sampling is not consistent!
Why?
Original: you observe the evidence.
Now: you set the evidence.
The sampling distribution has changed!



Likelihood Weighting

- Consider $P(shape|color = blue)$



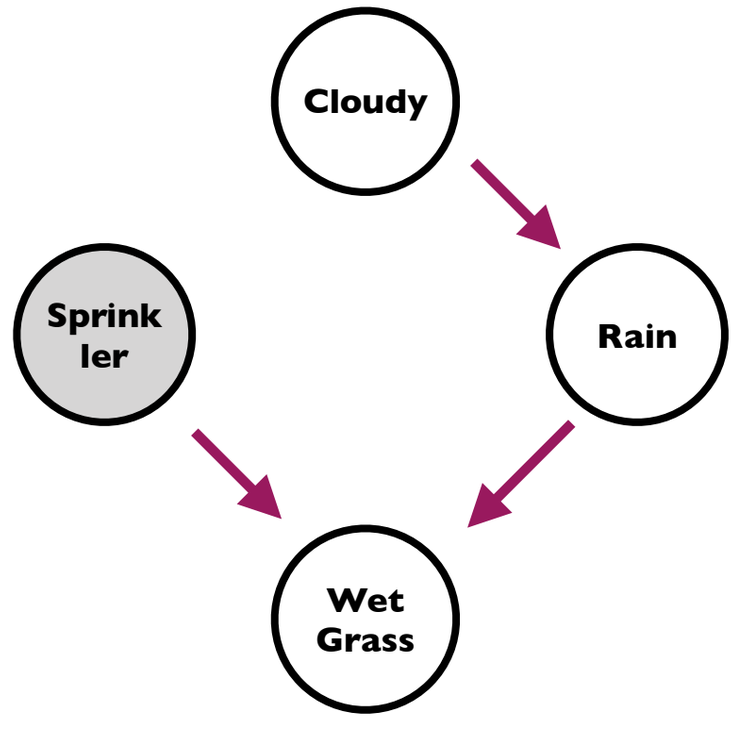
- Rejection sampling wastes lots of samples:

~~pyramid, green~~
~~pyramid, red~~
sphere, blue
~~cube, red~~
~~sphere, green~~

pyramid, blue
pyramid, blue
sphere, blue
cube, blue
sphere, blue

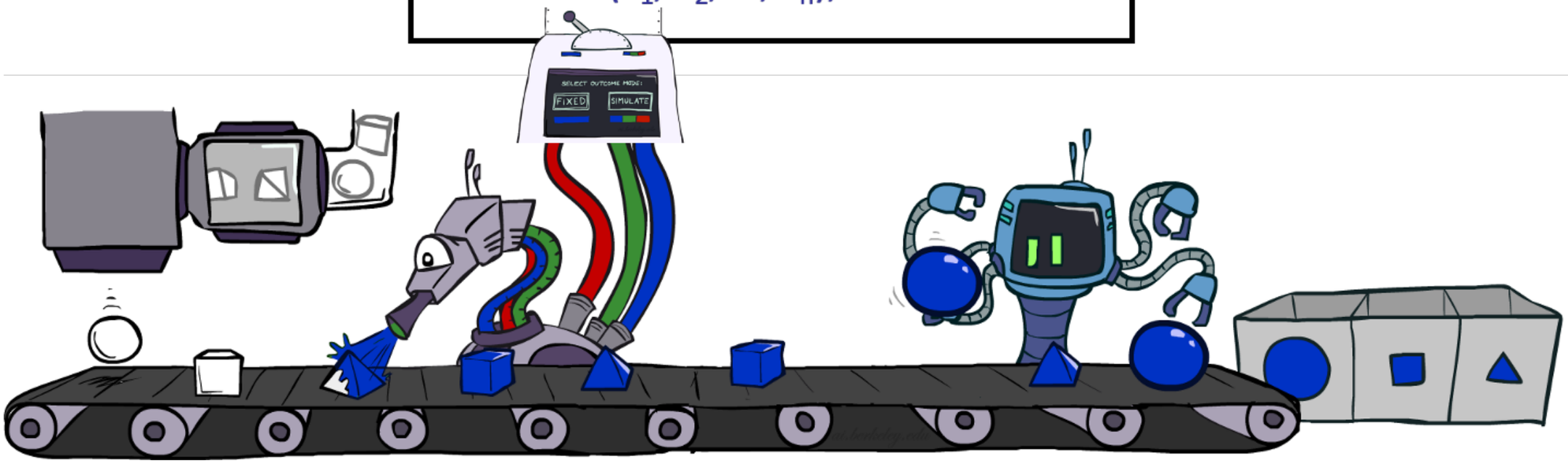
- Why not generate samples directly with conditioning on the evidence?

The sampling is not consistent!
Why?
Original: you observe the evidence.
Now: you set the evidence.
The sampling distribution has changed!



Likelihood Weighting

- IN: evidence instantiation
- w = 1.0
- for i=1, 2, …, n
  - if X_i is an evidence variable
    - X_i = observation x_i for X_i
    - Set w = w * $P(x_i \mid Parents(X_i))$
  - else
    - Sample x_i from $P(X_i \mid Parents(X_i))$
- return $(x_1, x_2, \ldots, x_n)$, w

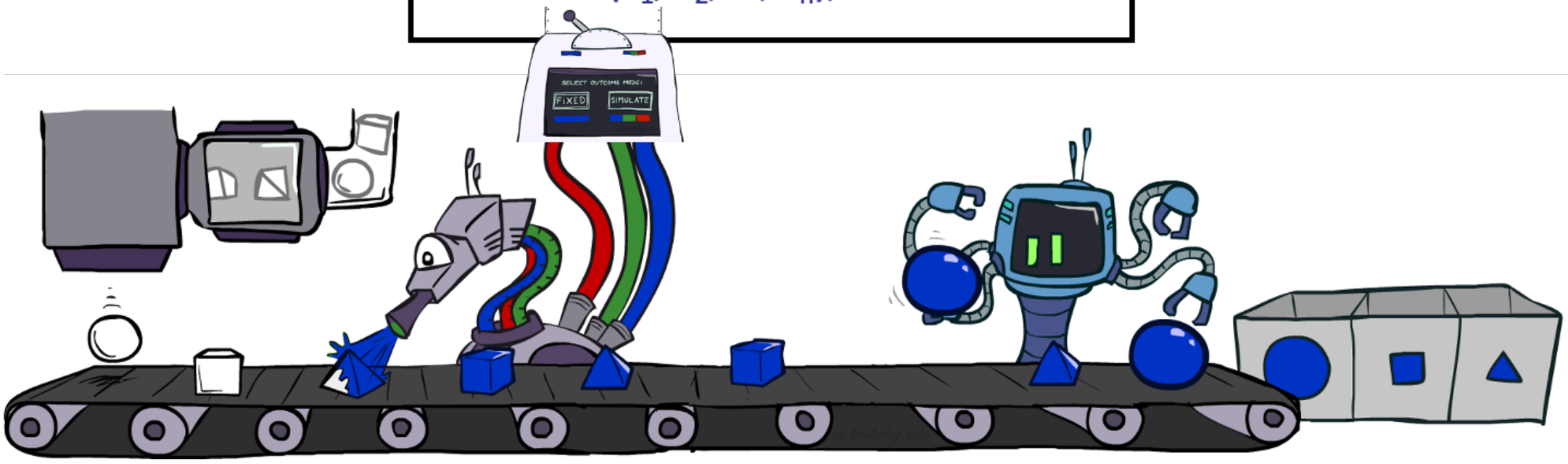


Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

- IN: evidence instantiation
- w = 1.0
- for i=1, 2, …, n
  - if X_i is an evidence variable
    - X_i = observation x_i for X_i
    - Set w = w * $P(x_i \mid Parents(X_i))$
  - else
    - Sample x_i from $P(X_i \mid Parents(X_i))$
- return $(x_1, x_2, \ldots, x_n)$, w

Assign weights to evidence! the idea of importance weighting

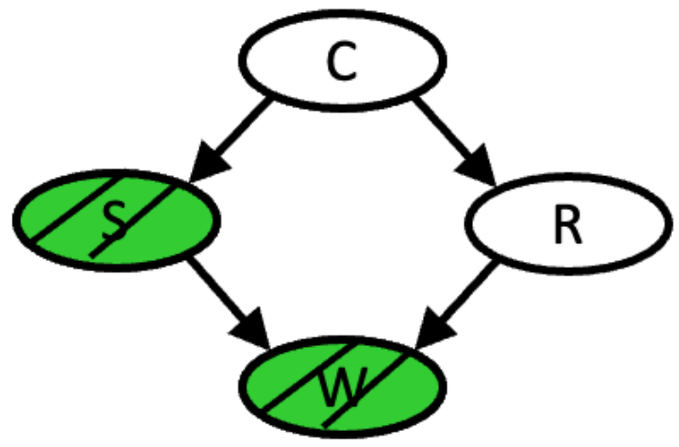


Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

- Sampling distribution if z sampled and e fixed evidence

$$S_{WS}(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{l} P(z_i | \text{Parents}(Z_i))$$



- Now, samples have weights

$$w(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{m} P(e_i | \text{Parents}(E_i))$$

- Together, weighted sampling distribution is consistent

$$S_{\text{WS}}(z, e) \cdot w(z, e) = \prod_{i=1}^{l} P(z_i | \text{Parents}(z_i)) \prod_{i=1}^{m} P(e_i | \text{Parents}(e_i))$$

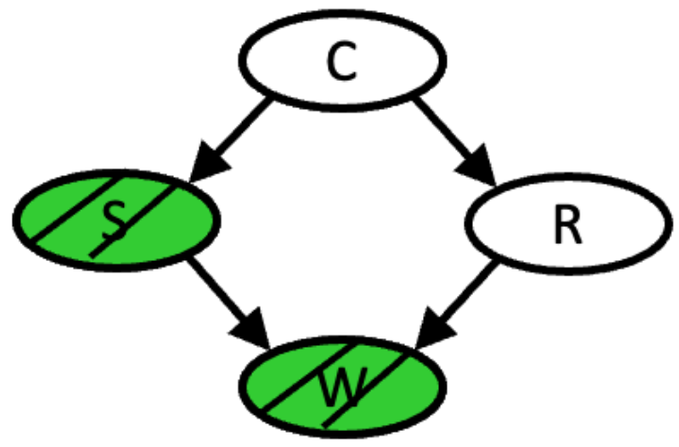
$$= P(\mathbf{z}, \mathbf{e})$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

- Sampling distribution if z sampled and e fixed evidence

$$S_{WS}(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{l} P(z_i | \text{Parents}(Z_i))$$



- Now, samples have weights

$$w(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^{m} P(e_i | \text{Parents}(E_i))$$

- Together, weighted sampling distribution is consistent

$$S_{\text{WS}}(z, e) \cdot w(z, e) = \prod_{i=1}^{l} P(z_i | \text{Parents}(z_i)) \prod_{i=1}^{m} P(e_i | \text{Parents}(e_i))$$

$$= P(\mathbf{z}, \mathbf{e})$$

And then do normalization to prob.

Slide courtesy: Dan Klein & Pieter Abbeel

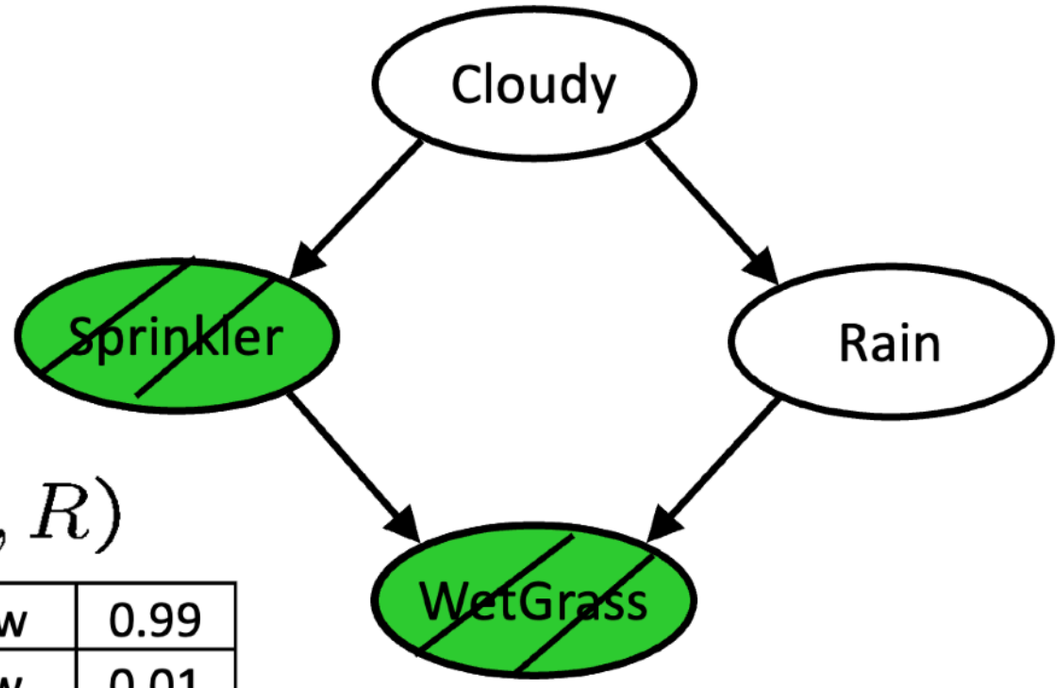
Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

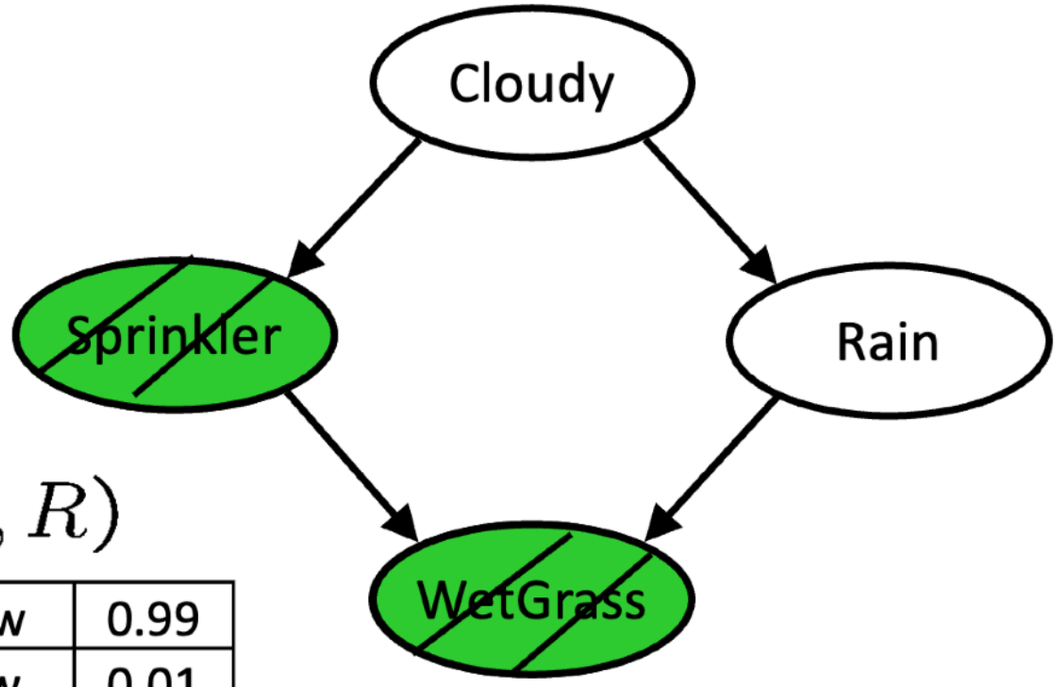
$$w = 1.0$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

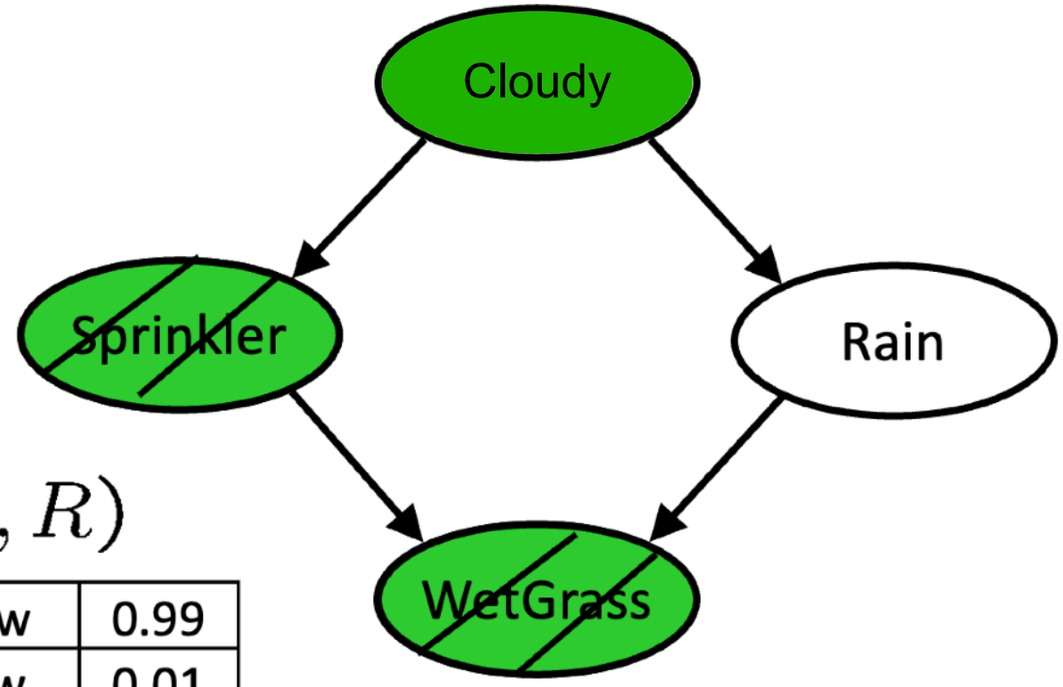
$$w = 1.0$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

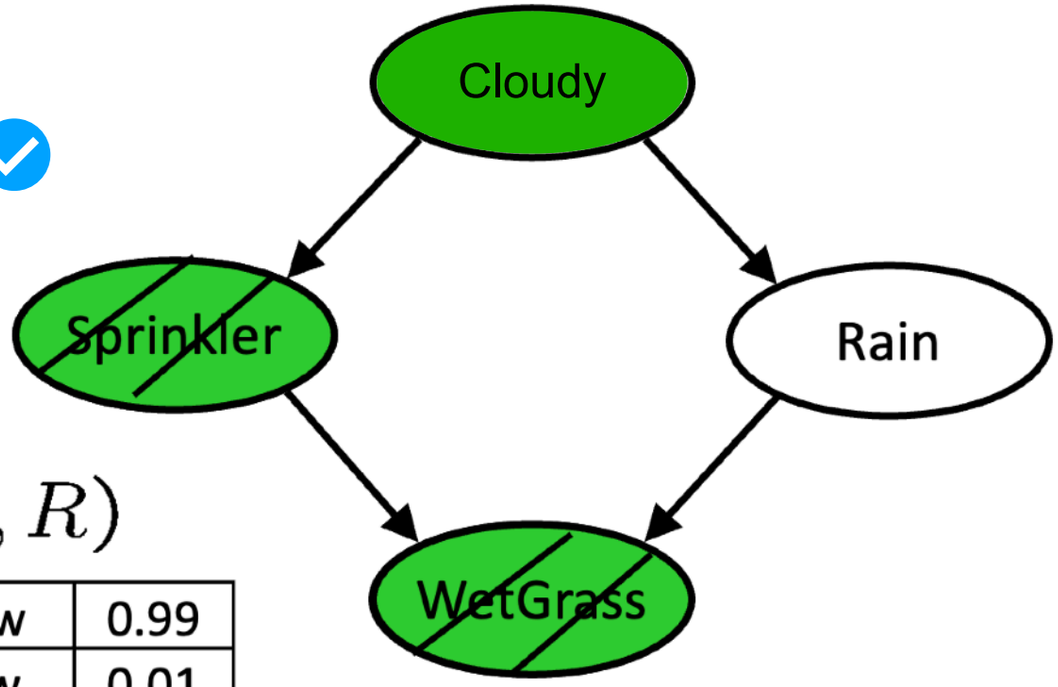
$$w = 1.0$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

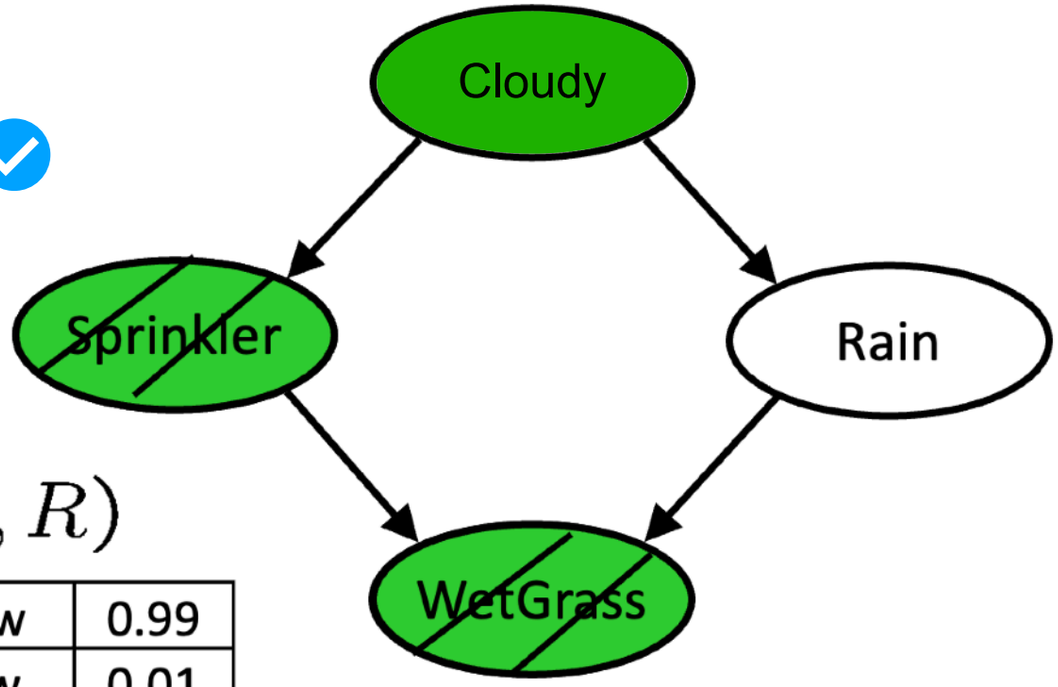
$$w = 1.0$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

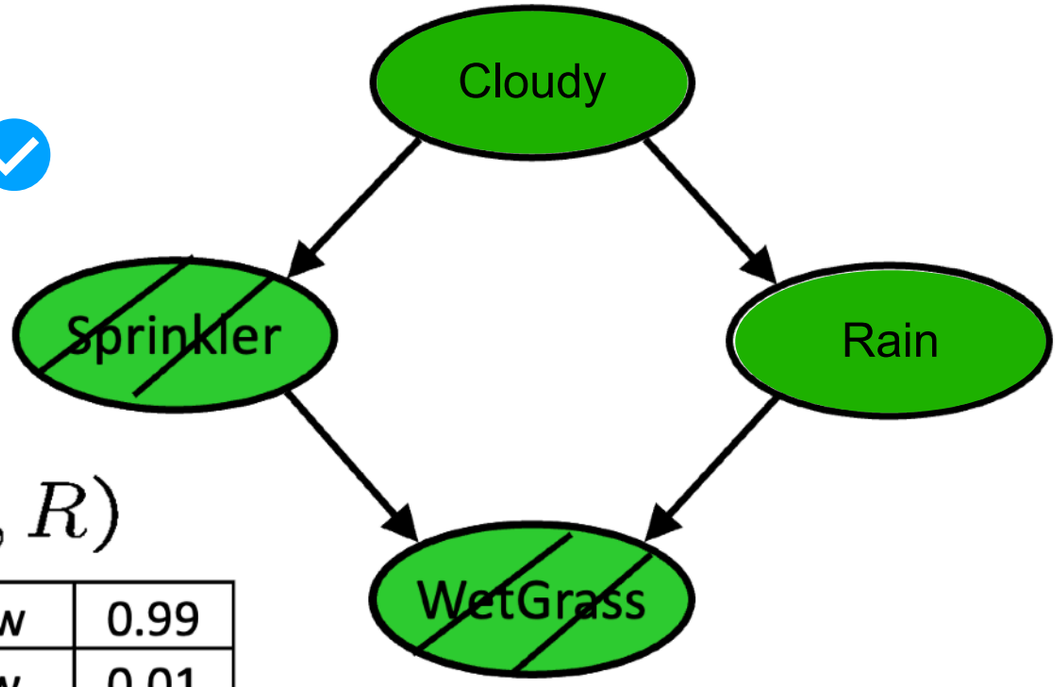
$$w = 1.0 \times 0.1$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

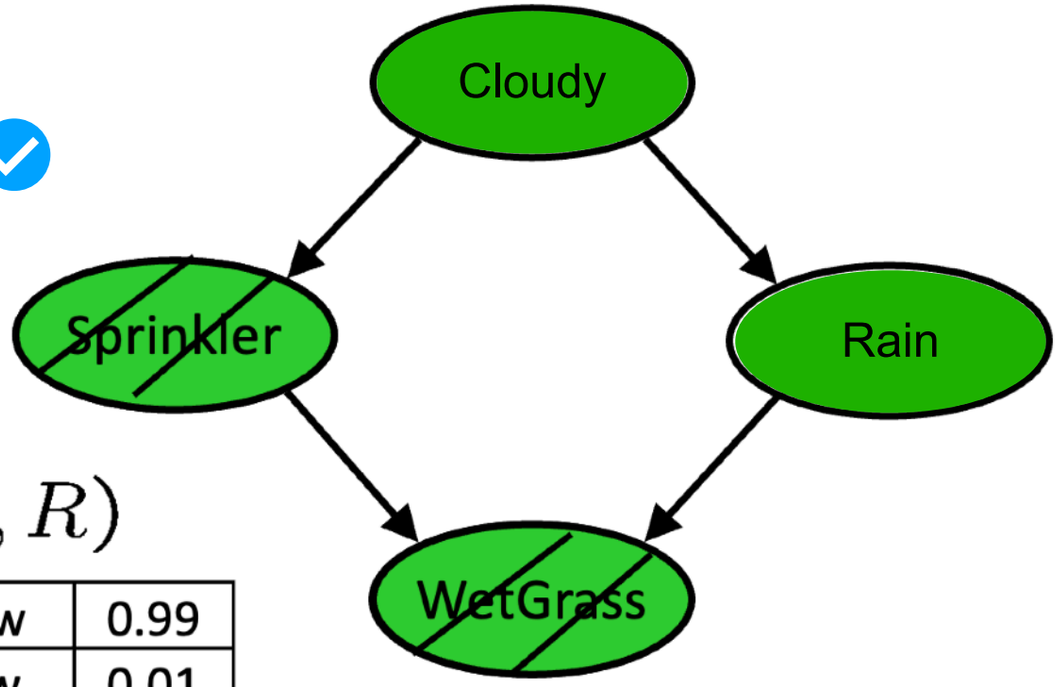
$$w = 1.0 \times 0.1$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

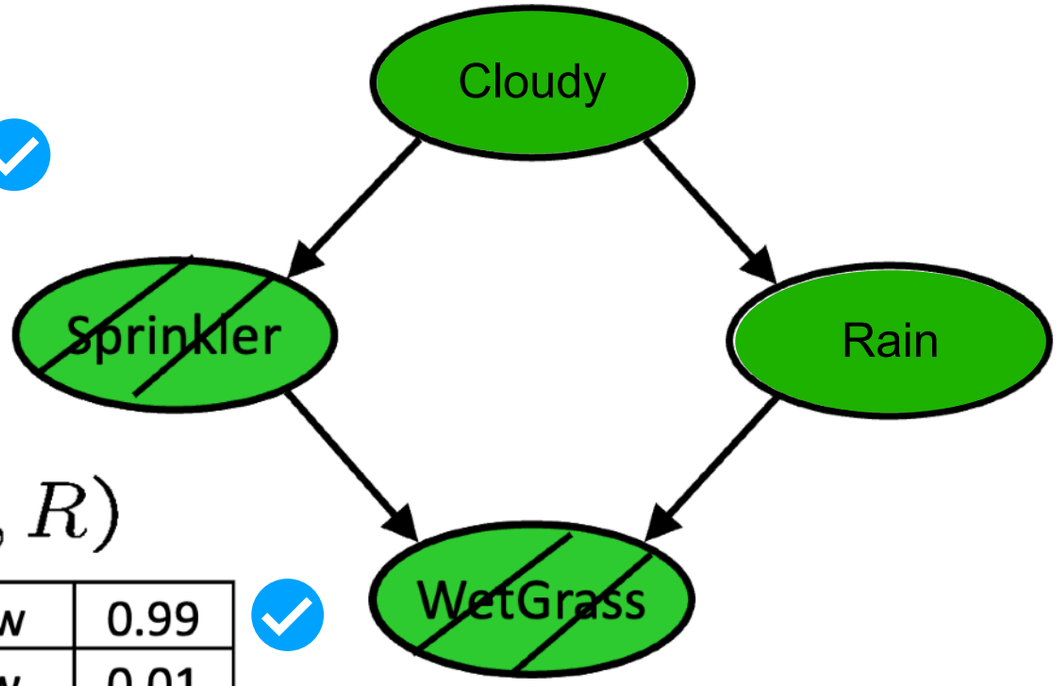
$$w = 1.0 \times 0.1$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

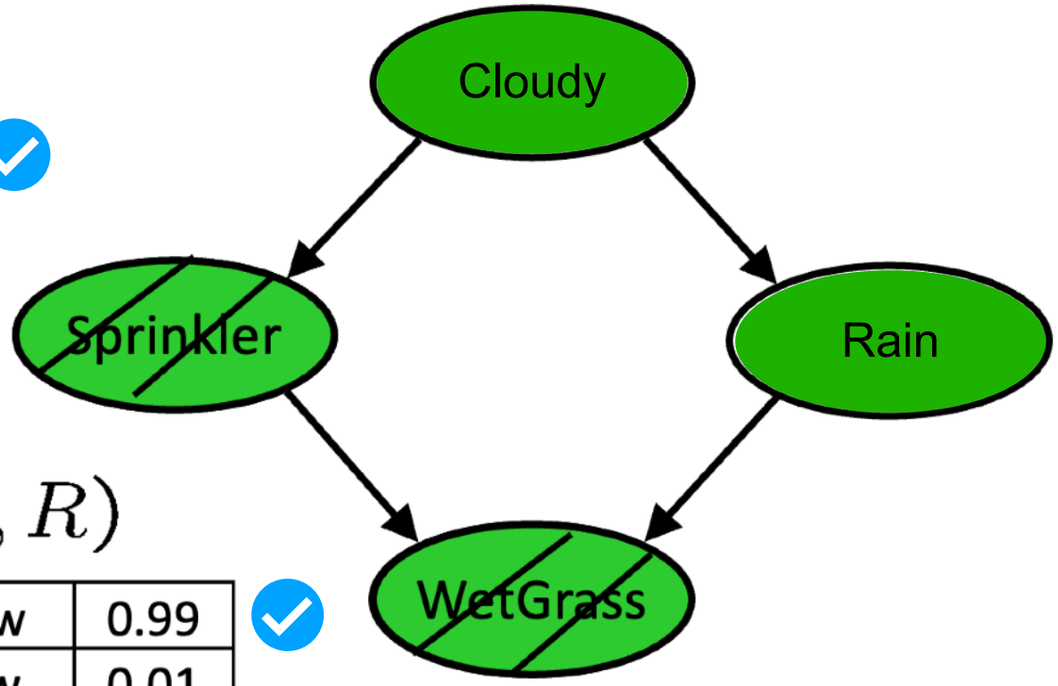
$$w = 1.0 \times 0.1$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

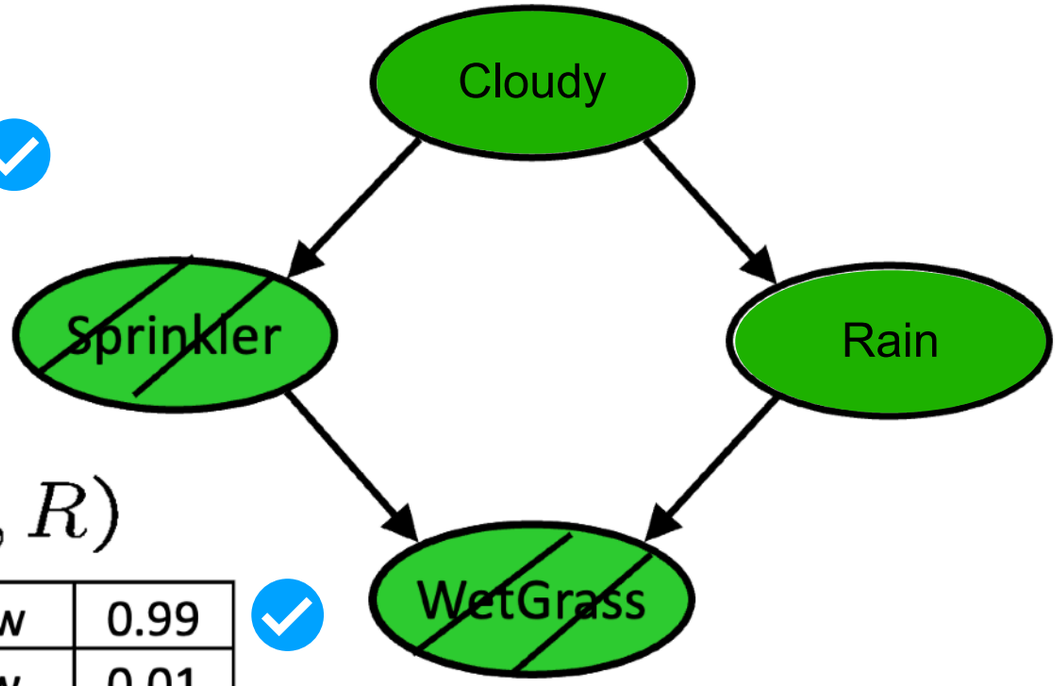
$$w = 1.0 \times 0.1 \times 0.99$$

Slide courtesy: Dan Klein & Pieter Abbeel

Likelihood Weighting

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

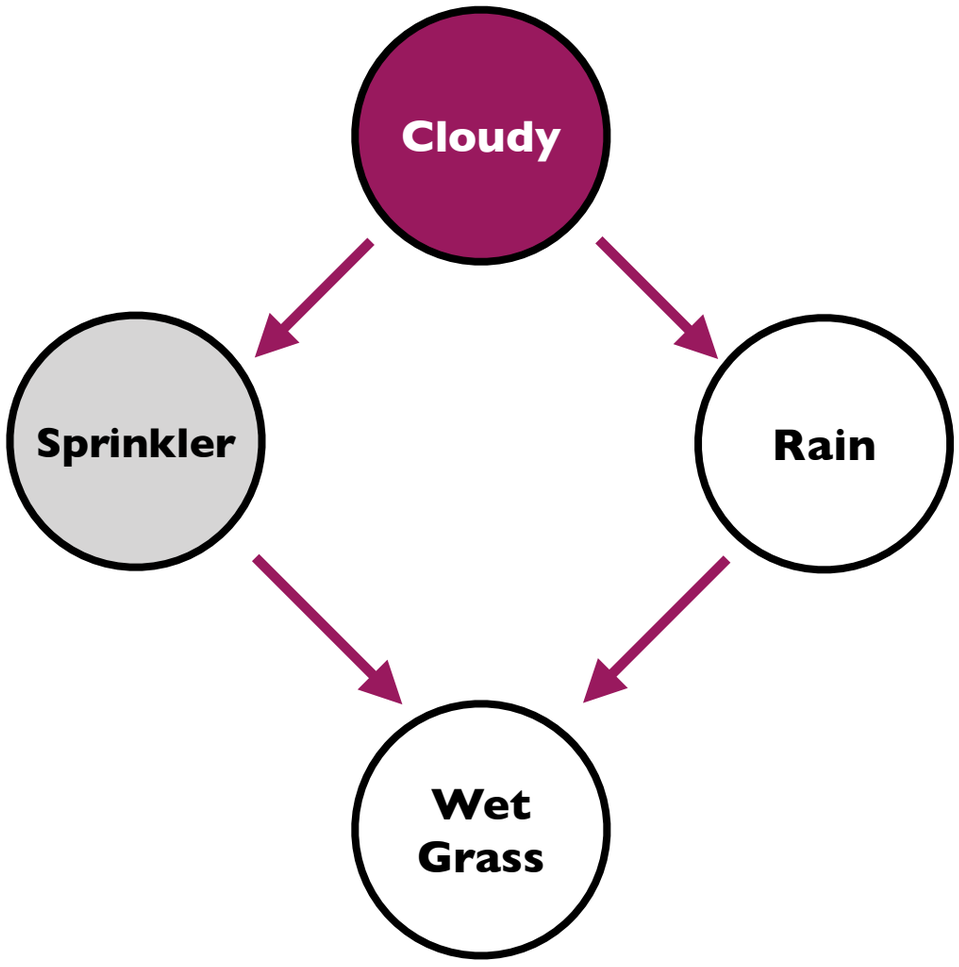
+c, +s, +r, +w

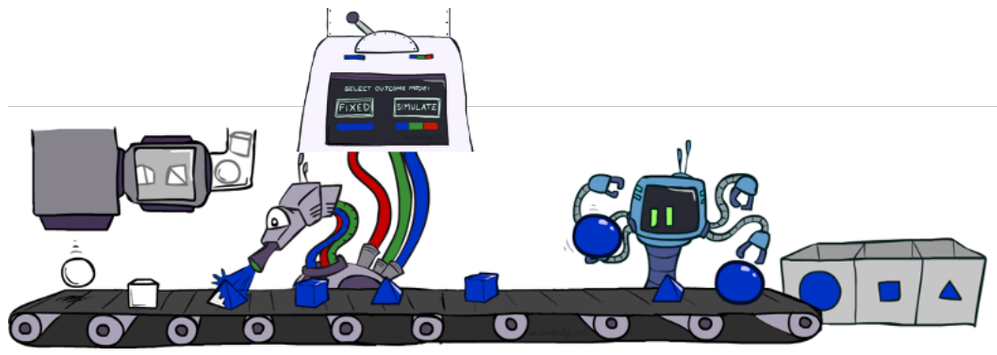
$$w = 1.0 \times 0.1 \times 0.99$$

Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

- In likelihood weighting, the evidence variables influence downstream (children) variables, but not upstream (parent) variables!





Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

- *Procedure:* keep track of a full instantiation $x_1, x_2, \ldots, x_n$. Start with an arbitrary instantiation consistent with the evidence. Sample one variable at a time, conditioned on all the rest, but keep evidence fixed. Keep repeating this for a long time.

- *Property:* in the limit of repeating this infinitely many times the resulting sample is coming from the correct distribution

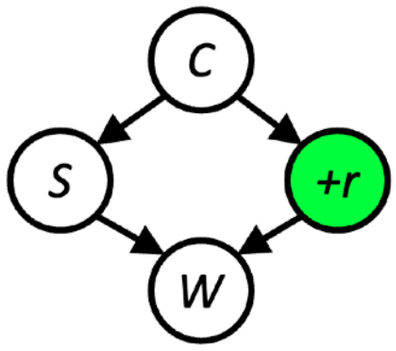
- *Rationale*: both upstream and downstream variables condition on evidence.

- In contrast: likelihood weighting only conditions on upstream evidence, and hence weights obtained in likelihood weighting can sometimes be very small. Sum of weights over all samples is indicative of how many "effective" samples were obtained, so want high weight.

Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

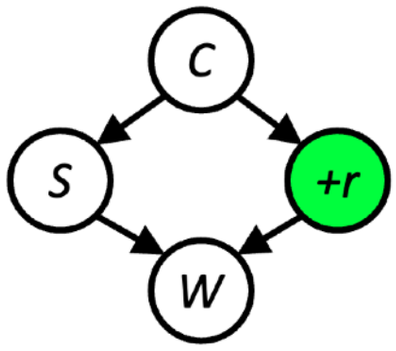
- Step 1: Fix evidence
  - R = +r



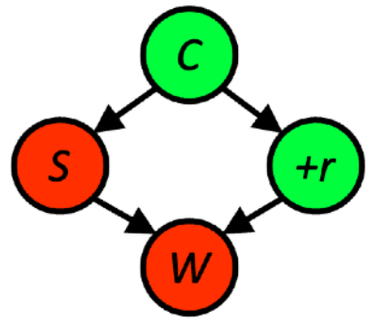
Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

- Step 1: Fix evidence
  - R = +r



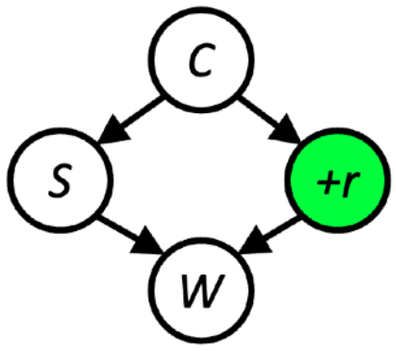
- Step 2: Initialize other variables
  - Randomly



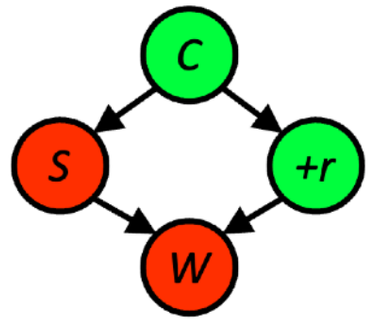
Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

- Step 1: Fix evidence
  - R = +r



- Step 2: Initialize other variables
  - Randomly

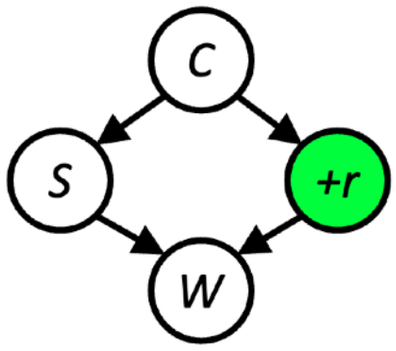


- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from P(X | all other variables)

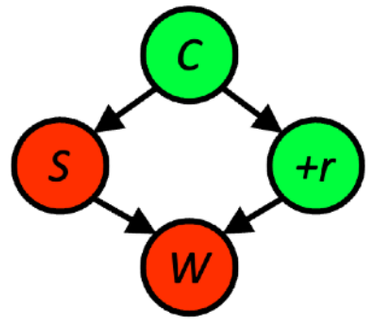
Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

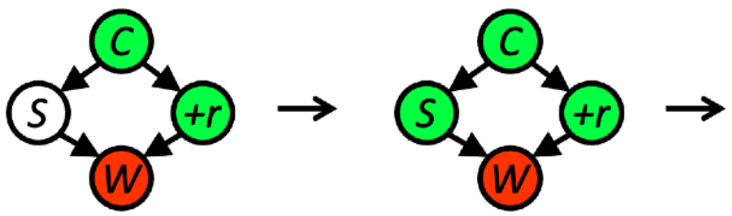
- Step 1: Fix evidence
  - R = +r



- Step 2: Initialize other variables
  - Randomly



- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from P(X | all other variables)

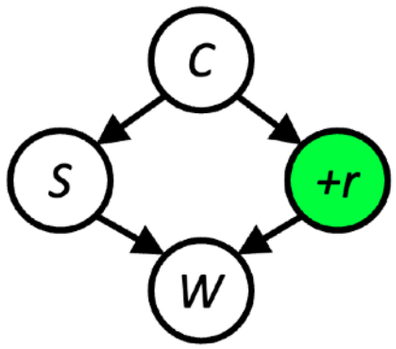


Sample from $P(S|+c,-w,+r)$

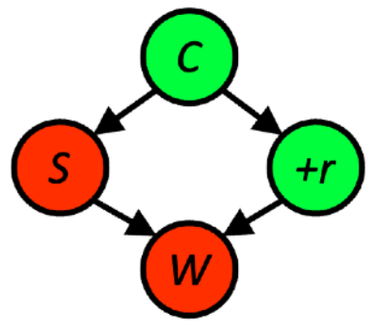
Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

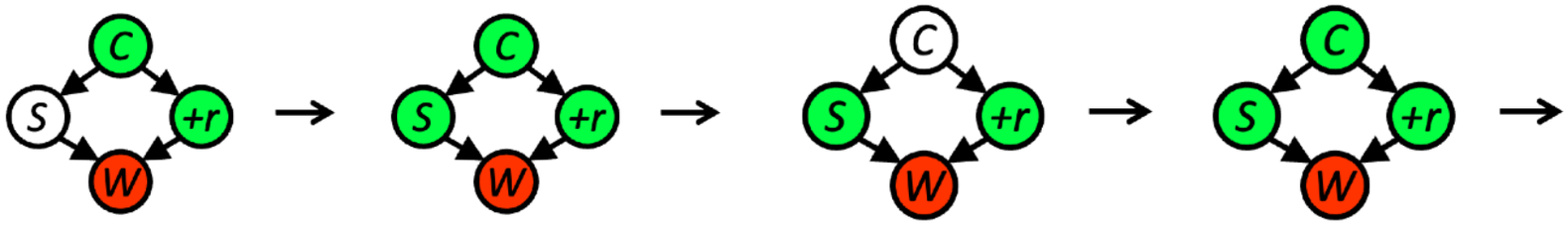
- Step 1: Fix evidence
  - R = +r



- Step 2: Initialize other variables
  - Randomly



- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from P(X | all other variables)

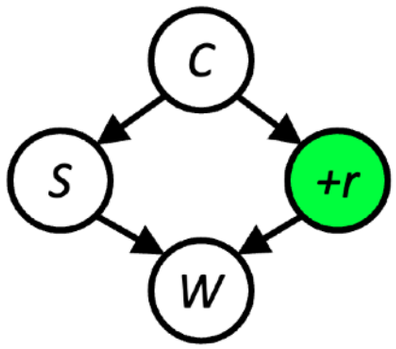


Sample from $P(S|+c,-w,+r)$ Sample from $P(C|+s,-w,+r)$

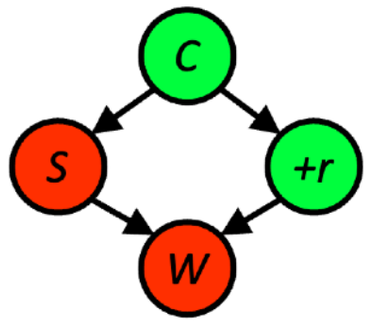
Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

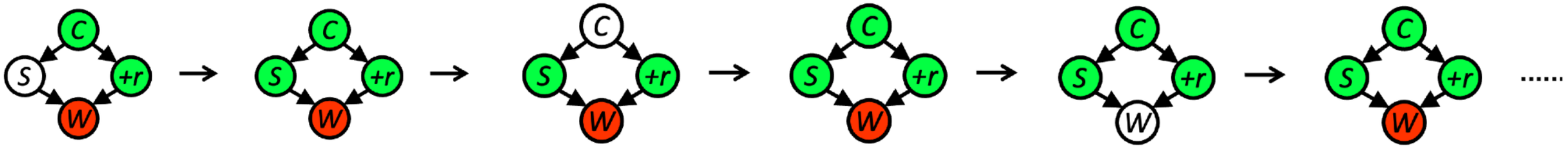
- Step 1: Fix evidence
  - R = +r



- Step 2: Initialize other variables
  - Randomly



- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from P(X | all other variables)

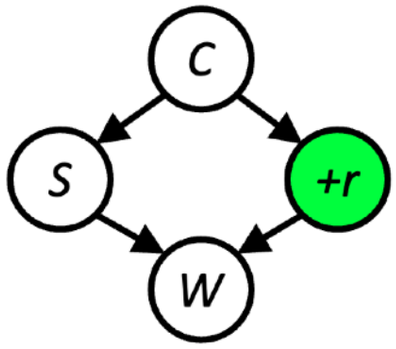


Sample from $P(S|+c,-w,+r)$ Sample from $P(C|+s,-w,+r)$ Sample from $P(W|+s,+c,+r)$

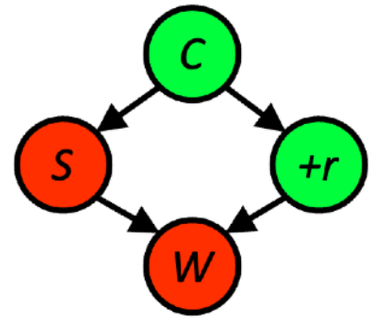
Slide courtesy: Dan Klein & Pieter Abbeel

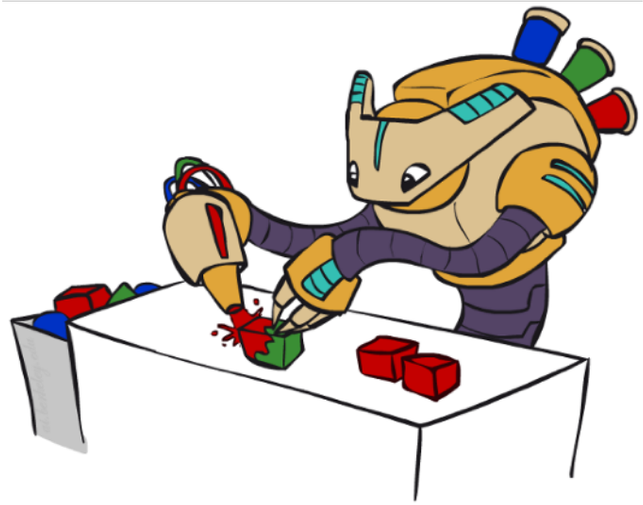
Gibbs Sampling

- Step 1: Fix evidence
  - R = +r

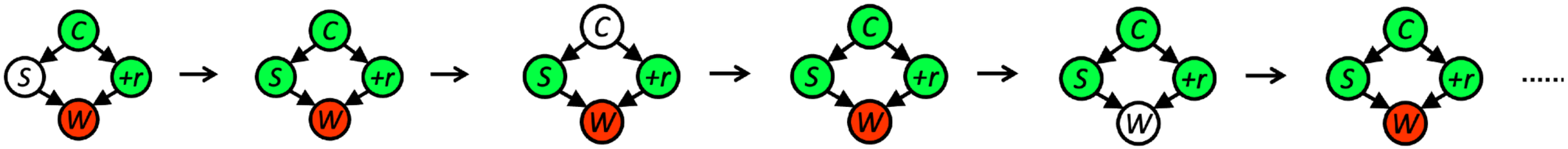


- Step 2: Initialize other variables
  - Randomly





- Steps 3: Repeat
  - Choose a non-evidence variable X
  - Resample X from P(X | all other variables)

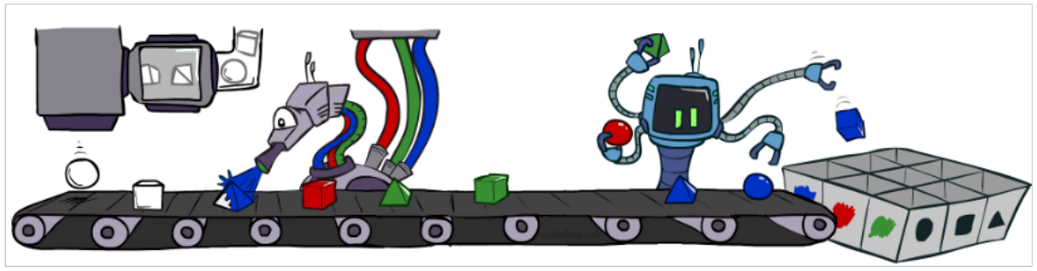


Sample from $P(S|+c,-w,+r)$ Sample from $P(C|+s,-w,+r)$ Sample from $P(W|+s,+c,+r)$

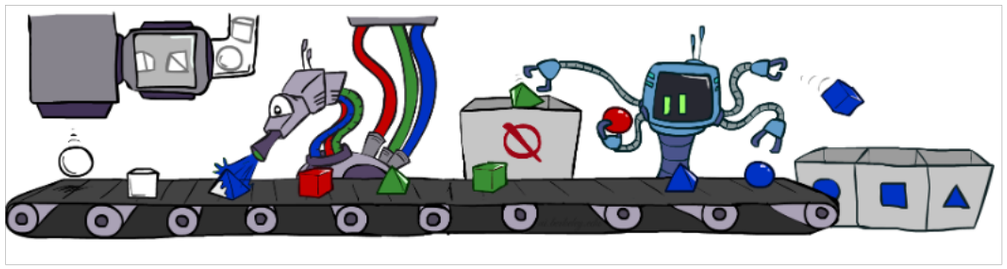
Slide courtesy: Dan Klein & Pieter Abbeel

Gibbs Sampling

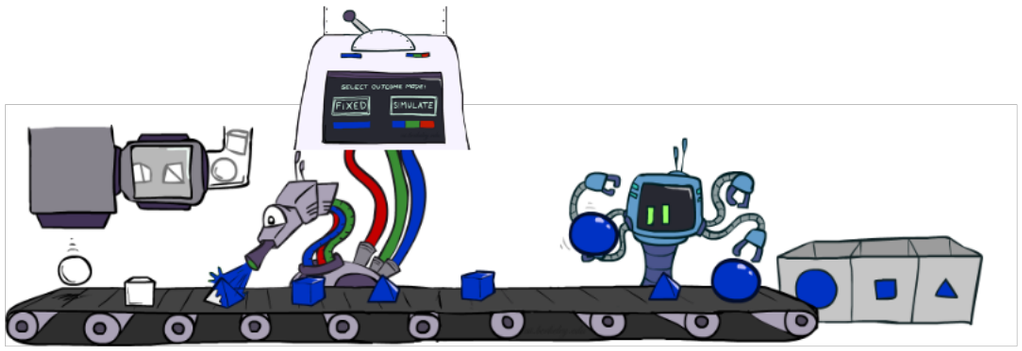
- Prior Sampling P



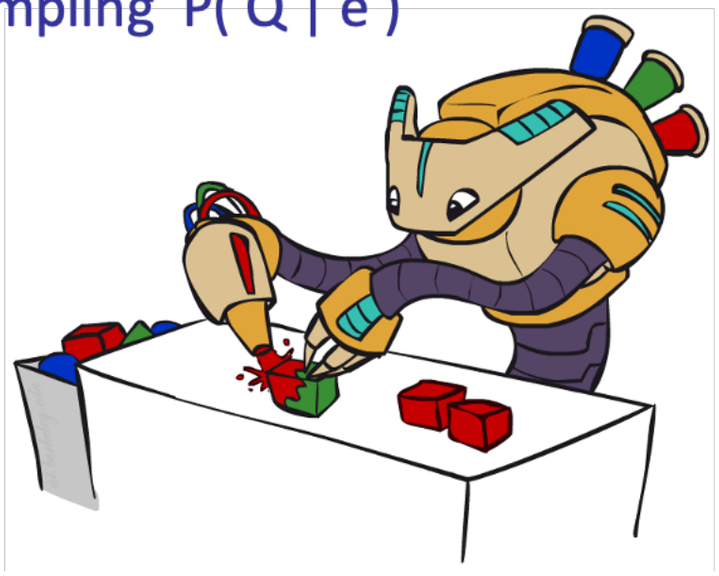
- Rejection Sampling P(Q | e)



- Likelihood Weighting P(Q | e)



- Gibbs Sampling P(Q | e)



Slide courtesy: Dan Klein & Pieter Abbeel

Knowledge Reasoning: III

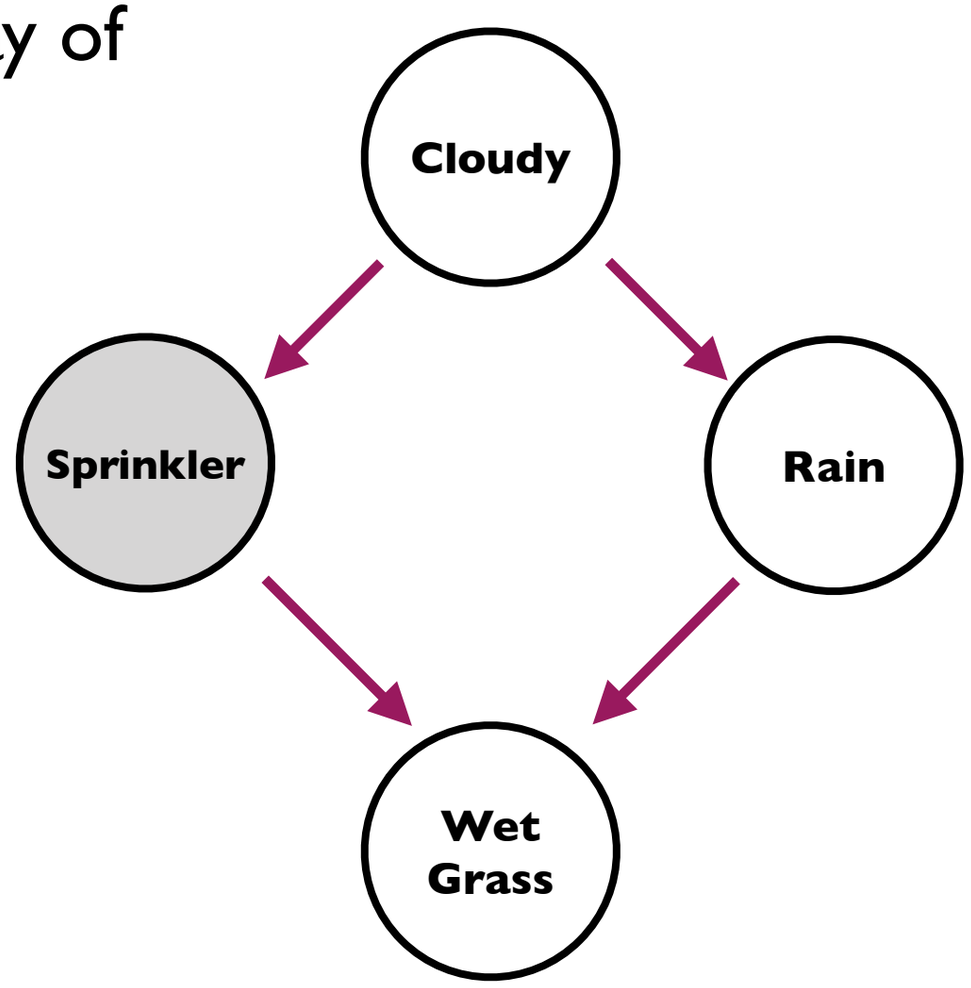
- Probabilistic Reasoning
  - Bayes net: Approximate inference
- Causal reasoning
- Take-home messages

Causation vs. Correlation

- Bayesian networks encode joint distributions.
- Joint distributions can be factored in different ways.
- Arrows in BNs only determine one way of factoring.

The directions of correlations can be represented in many ways.

The directions of causation is unique!



Why Causal Relationship is Important for AI?

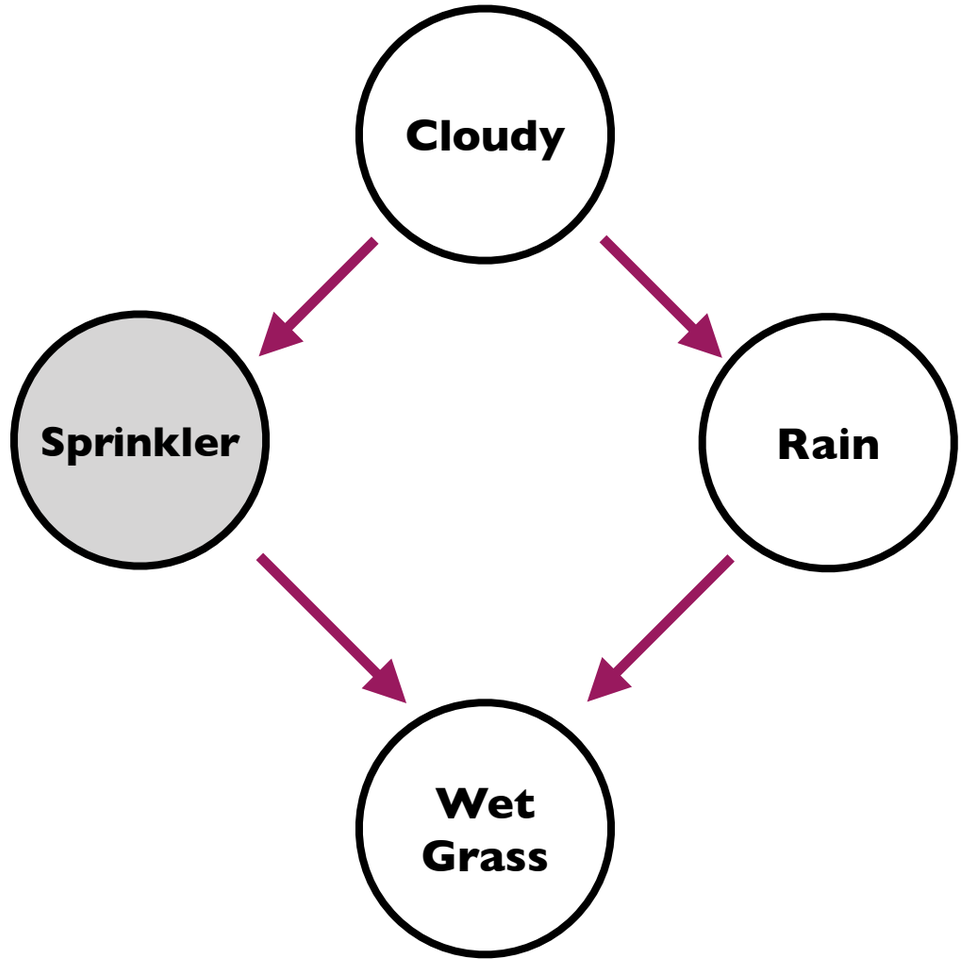
The causal knowledge is robust against environmental changes

- Knowing whether the grass is wet changes the conditional probability

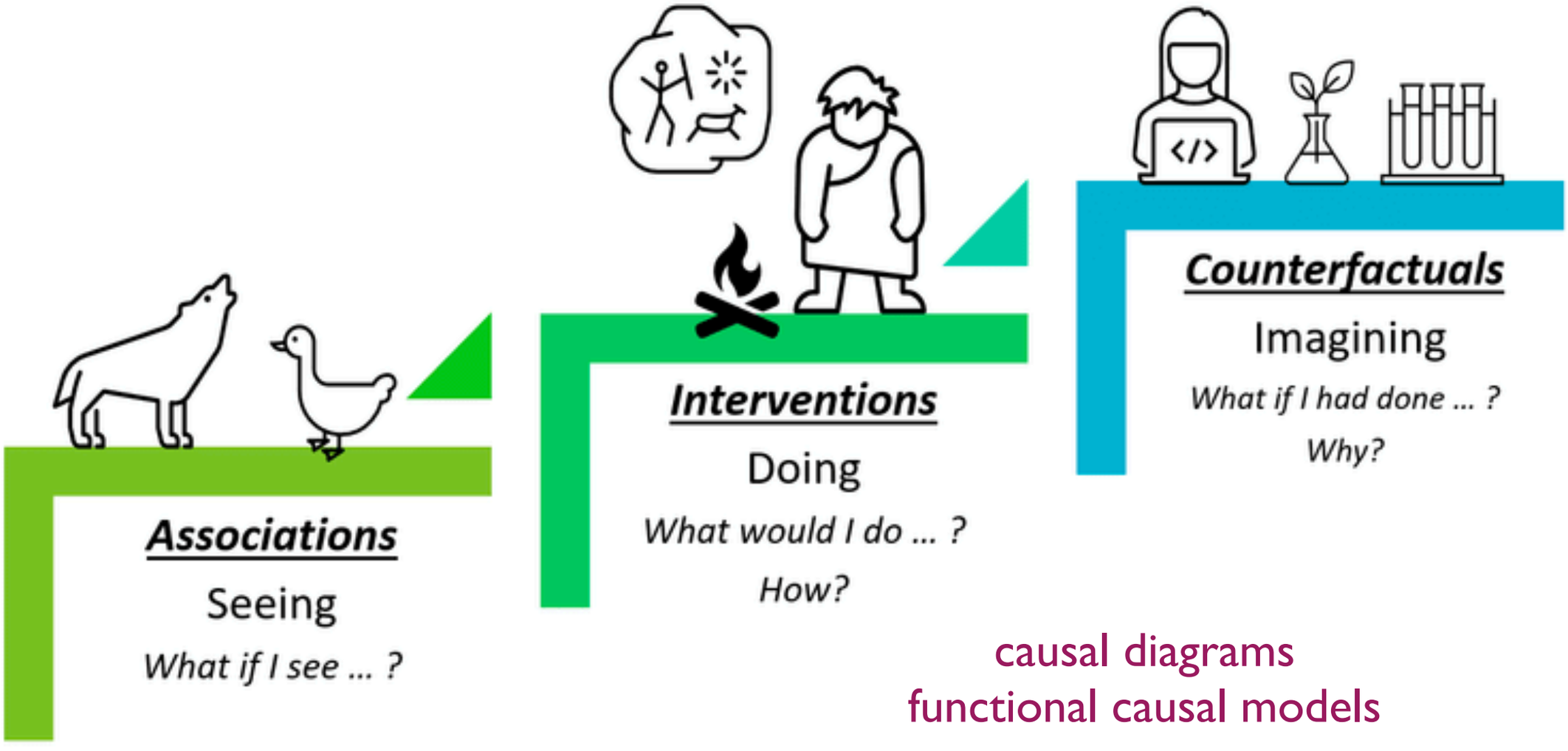
$$P(rain|sprinkler, cloudy)$$

$$P(rain|sprinkler, cloudy, grass = wet)$$

- But the causal relationship among sprinkler, cloudy, and rain should not change!



The Ladder of Causality



joint distributions
like BNs

causal diagrams
functional causal models

Simpson's Paradox

Treatment / Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Which treatment is better? Why?

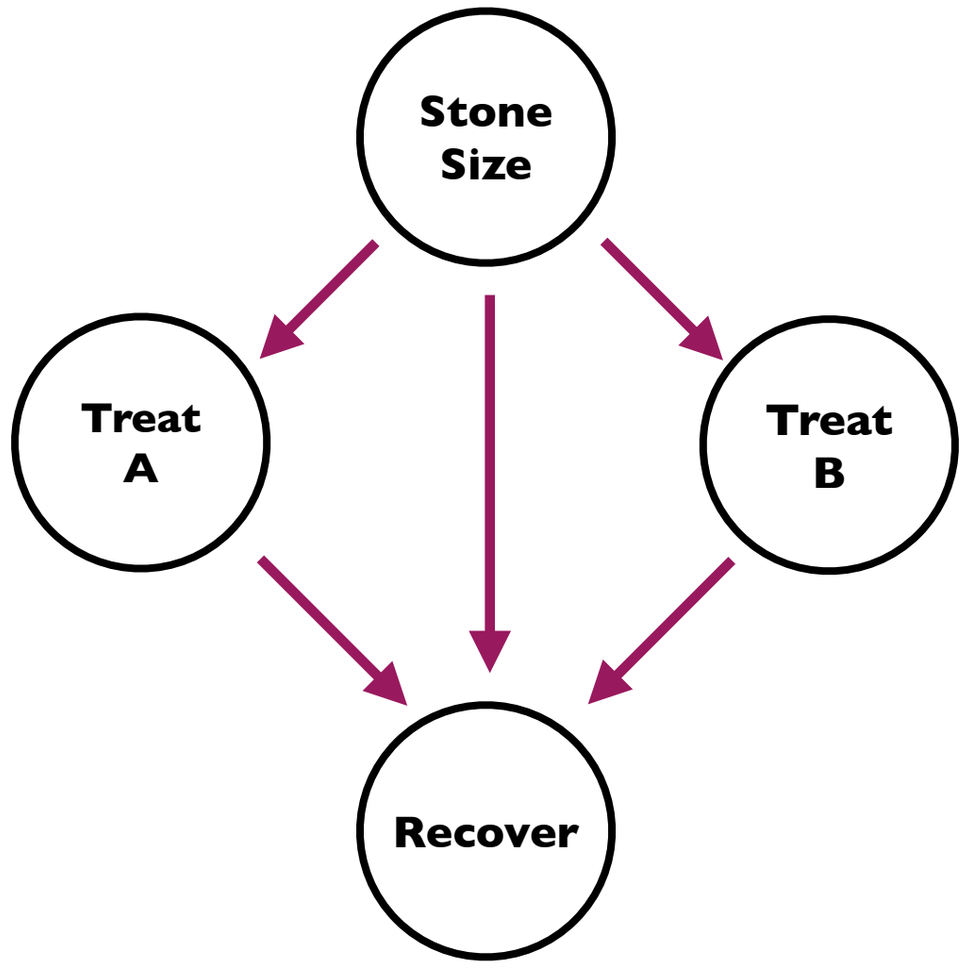
Simpson's Paradox

Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Which treatment is better? Why?

Large stones are harder, and treatment B is cheaper

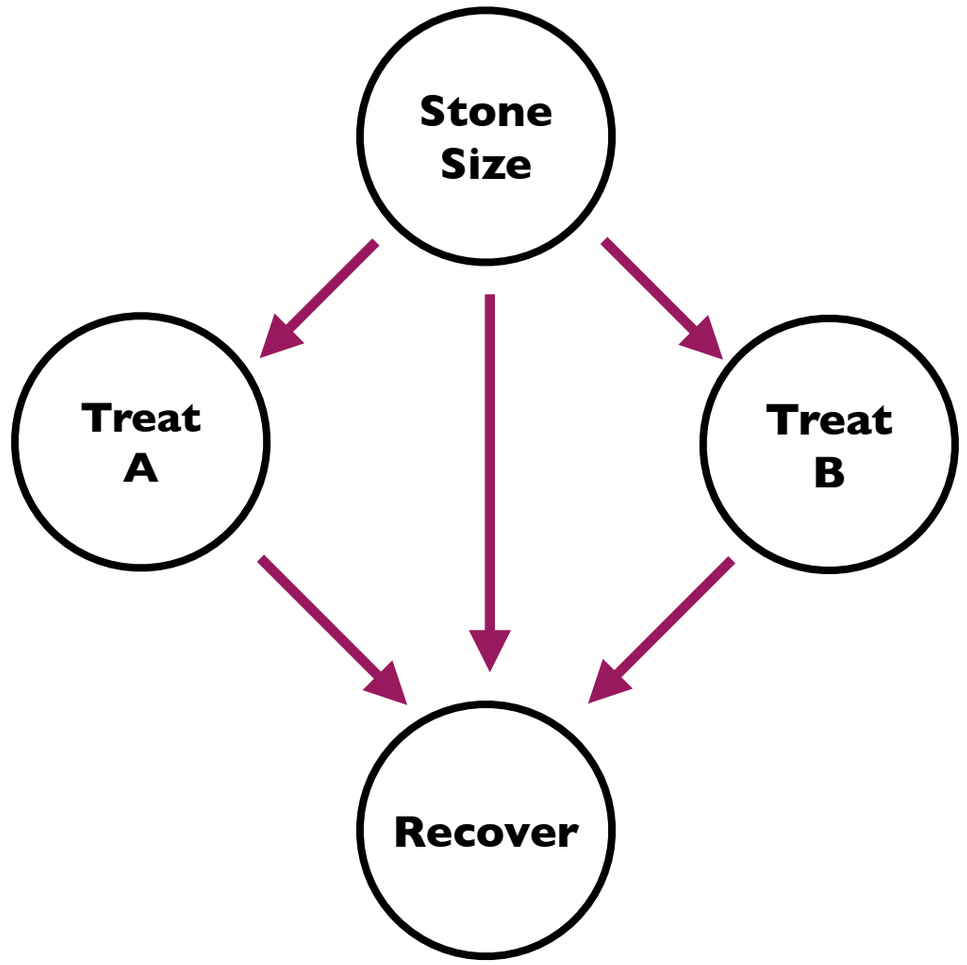
Simpson's Paradox



- Similar example: air conditioner on vs. feeling hot

Discovering causal relationship should block those underlying effects on the causes!

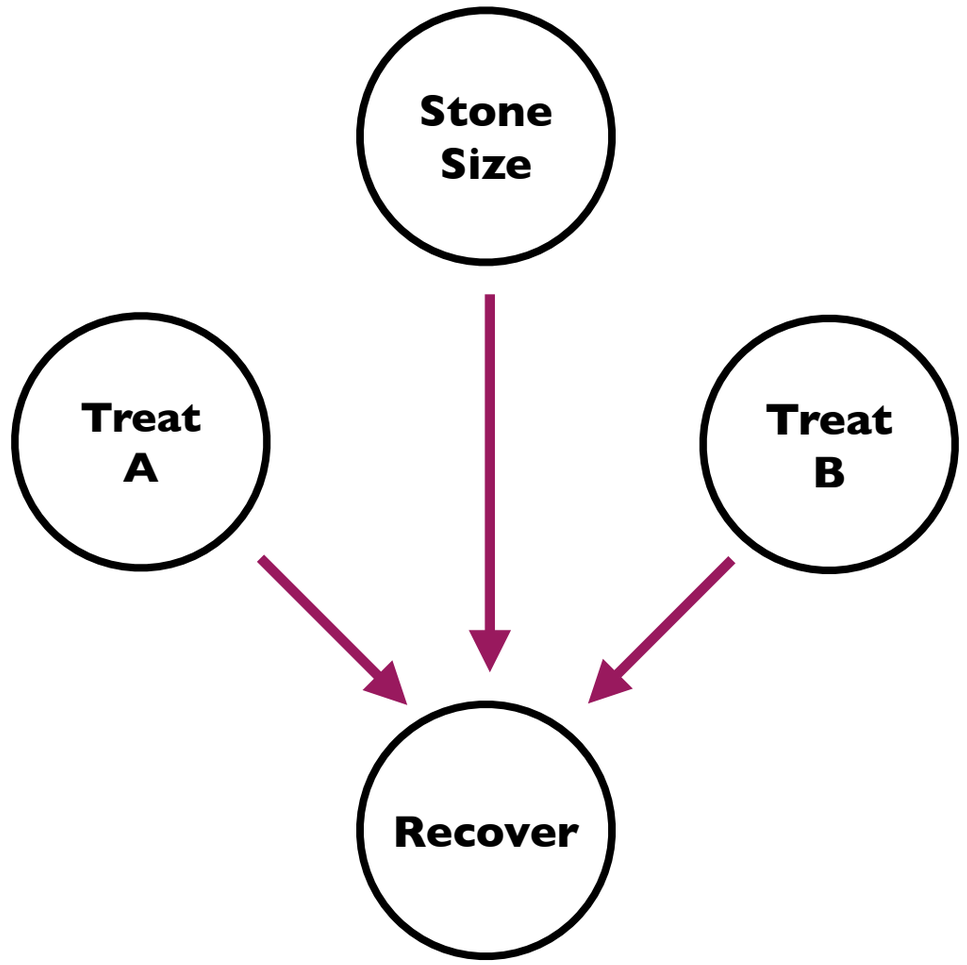
Intervention



- The key idea is to consider the intervention $P(recover|do(treatA))$ instead of the association $P(recover|treatA)$

- Common method: random controlled experiments!

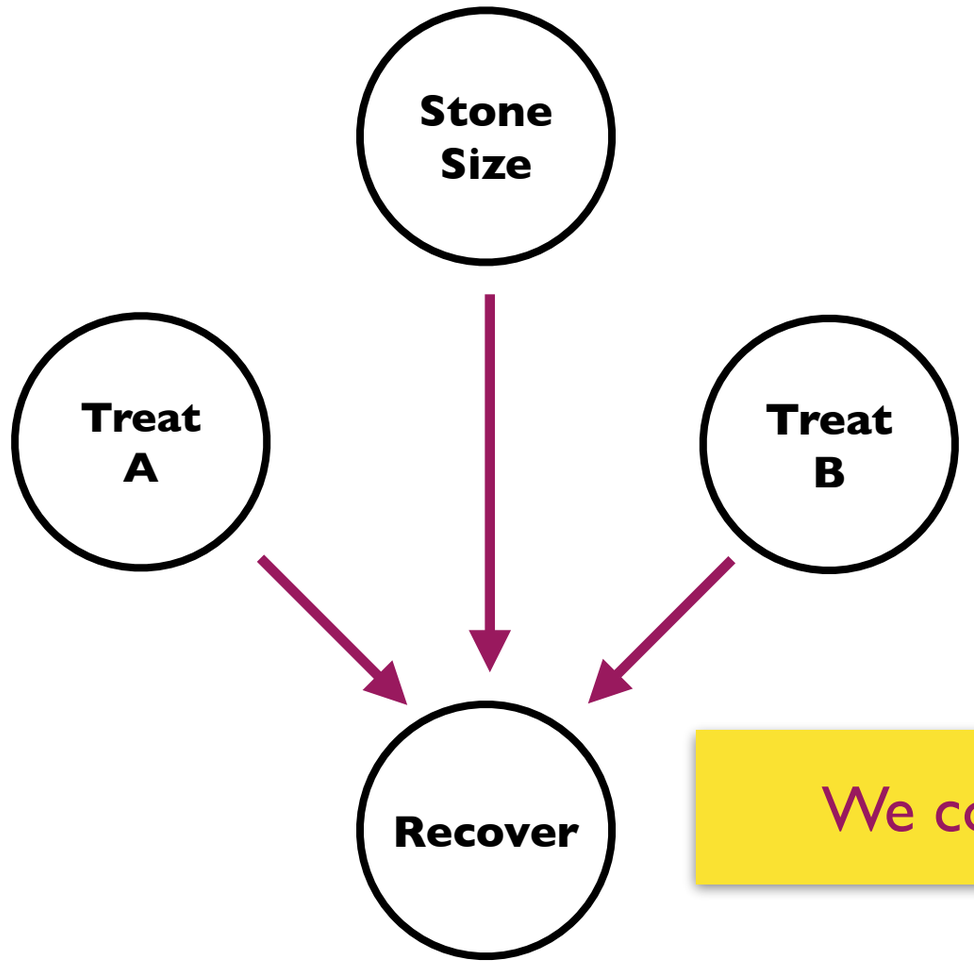
Intervention



- The key idea is to consider the intervention $P(recover|do(treatA))$ instead of the association $P(recover|treatA)$

- Common method: random controlled experiments!

Intervention



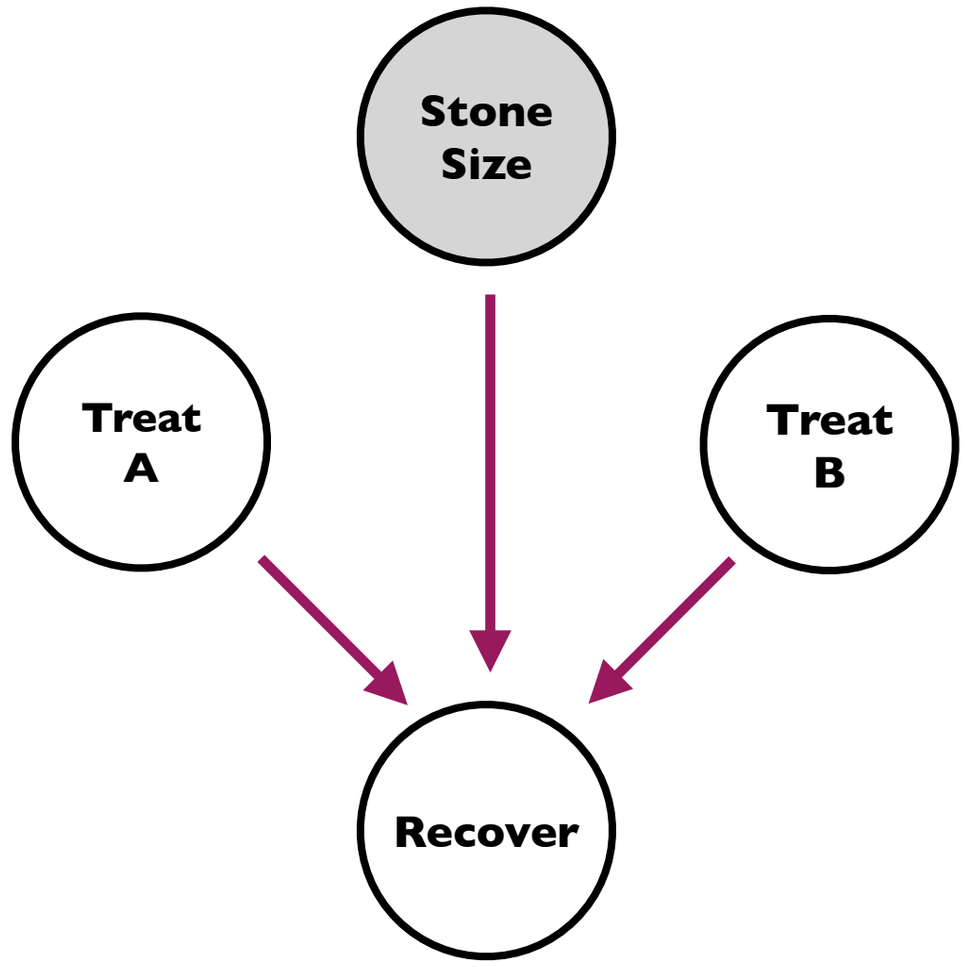
We consider a changed BN!

- The key idea is to consider the intervention $P(recover|do(treatA))$ instead of the association $P(recover|treatA)$

- Common method: random controlled experiments!

Back-Door Criterion

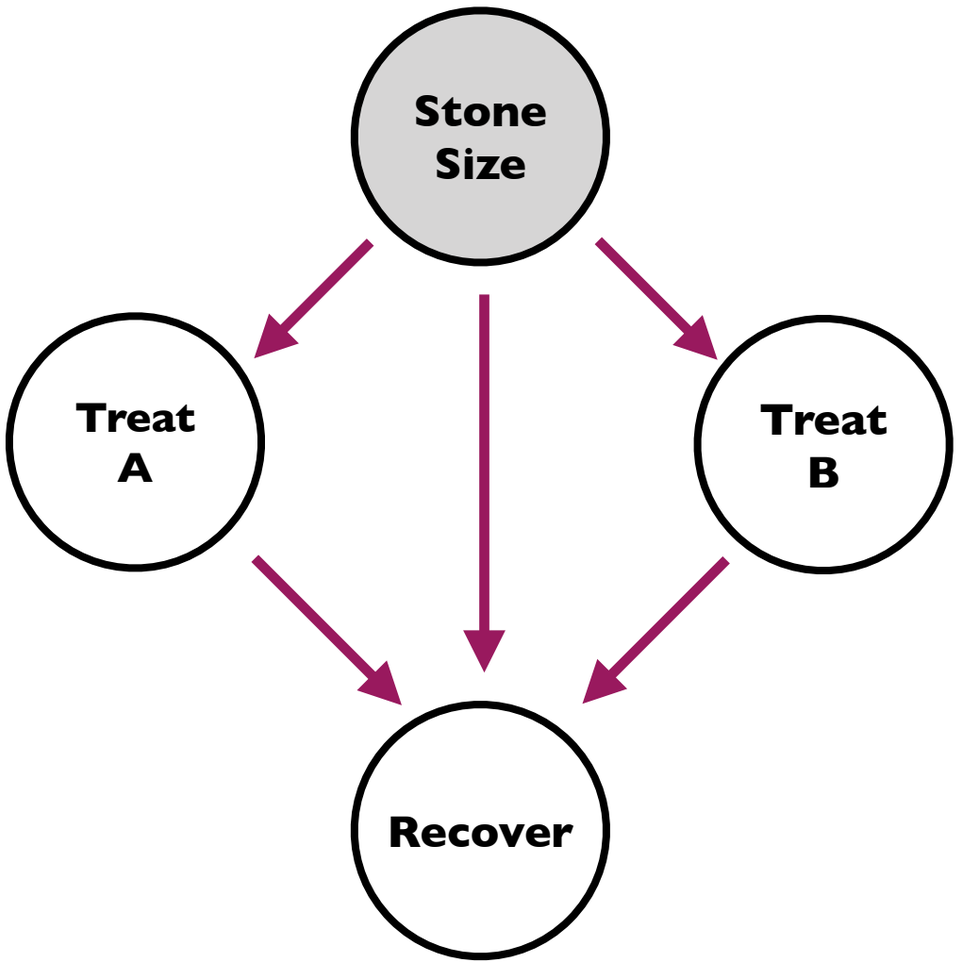
Treatment / Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



- Experiments are not always necessary. Can infer from observations!
- Just close the "back doors" by conditioning on parent variables.
- Many interesting algorithms.

Back-Door Criterion

Treatment / Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



- Experiments are not always necessary. Can infer from observations!
- Just close the "back doors" by conditioning on parent variables.
- Many interesting algorithms.

Counterfactuals

If the treatment was not given,
would the patient recover?

- We can not even get data to estimate!
- But they lie at heart of human intelligence.

Functional Causal Models

- We should know more than conditional probabilities: the underlying physical mechanism among causes and effects.

- Functional causal models:

$$x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n$$

x_i: effect; pa_i: control variables; u_i: unmodeled randomness

- Example: $x_i = \sum_{k \neq 1} \alpha_{ik} x_k + u_i, \quad i = 1, \ldots, n$

Counterfactuals

$$x = u_1,$$

$$y = xu_2 + (1 - x)(1 - u_2)$$

X: treatment
Y: death

Know: X=1, Y=1
Ask: whether
X=0, Y=0?

- Abduction: put the evidence into the equations:

$$u_1 = 1, u_2 = 1$$

- Action: set the new control variable:

$$x = 0$$

- prediction: get the new effect:

$$y = 0$$

Counterfactuals

$$x = u_1,$$

$$y = xu_2 + (1 - x)(1 - u_2)$$

X: treatment
Y: death

Know: X=1, Y=1
Ask: whether
X=0, Y=0?

- Abduction: put the evidence into the equations:

$$u_1 = 1, u_2 = 1$$

- Action: set the new control variable:

$$x = 0$$

- prediction: get the new effect:

$$y = 0$$

Similar to traveling in parallel universe

Knowledge Reasoning: III

- Probabilistic Reasoning
  - Bayes net: Approximate inference
- Causal reasoning
- Take-home messages

Take-Home Messages

- Approximate inference avoids the complexity of marginalization by assuming simpler distributions or sampling.

- More effectively using the evidences can significantly improve the efficiency of sampling.

- The ladder of causality: association, intervention, and counterfactual.

Thanks for your attention!
Discussions?

Acknowledgement: Many materials in this lecture are taken from
http://ai.berkeley.edu/lecture_slides.html